

University of Mannheim

College of Social Sciences – Department of Psychology



Replication of a Meta Analysis: A Crossroad Model

Diplomarbeit

Lehrstuhl Psychologie II

By Robin Würfel

Advisor: Prof. Dr. Michael Bosnjak

Second Advisor: Prof. Dr. Werner W. Wittmann

Mannheim, June 1st 2009

Address:

Mundenheimer Str. 241

67061 Ludwigshafen

Acknowledgments

I would like to thank everyone who provided support in the preparation of this thesis. In particular I would like to thank Holger Pahl for the fruitful teamwork and the hours of shared time, always making hard work appear like fun. Further more I would like to thank Prof. Dr. Michael Bosnjak for acting as the thesis advisor and his inspiration and guidance.

Additional thanks go to Sarah Hellmann, who served as a third coder and had to endure all 200 or so items to be coded during her own thesis phase. I would also like to thank Verena Würfel for her helpful comments on the earlier drafts of this thesis and especially Kasia Koczon, who corrected my paper with joy and always distracted and fed me when I needed it the most.

Mannheim, 6th June 2009

Robin Würfel

Index

1	Introduction	1
2	Theoretical Background	3
2.1	Meta Analysis – Development and a Five Step Approach	3
2.1.1	Problem Formulation	4
2.1.2	Data Collection.....	5
2.1.3	Evaluation of Data Points	5
2.1.4	Data Analysis and Interpretation.....	7
2.1.5	Presentation of Results	9
2.1.6	Different Meta Analyses Approaches.....	10
2.1.7	Common Critique Concerning Meta Analyses	11
2.2	Stayner (2007) – Lung Cancer Risk and Workplace Exposure to ETS.....	14
2.2.1	The Field of Research of “Environmental Tobacco Smoke (ETS)”	15
2.2.2	Stayner (2007) Lung Cancer Risk and Workplace Exposure to Environmental Tobacco Smoke – a Short Review.....	16
2.2.3	Main Results of Stayner et al. (2007).....	16
2.2.4	Flaws in Stayner et al. (2007).....	17
2.3	Subjectivity of Meta Analyses and Implementation of the Crossroad Model.....	22
2.3.1	Meta Analysis and its Subjectivity	23
2.4	A Crossroad Model.....	26
2.4.1	Finding the Upper and Lower Limits	28
2.4.2	About Missing Data	29
2.5	Advanced Analyses	29
2.5.1	Quality of Primary Studies.....	30
2.5.2	Correction of Study Artefacts: An Approach by Hunter and Schmidt	31
3	Study Questions.....	32
3.1	Replication Phases	32
3.2	Study Questions Related to the CRM.....	32
3.3	Study Questions Related to the Advanced Analysis	33
4	Methods.....	34

Index

4.1	Problem Formulation and Data Collection.....	34
4.2	Evaluation of Data Points	34
4.2.1	Basis for Data Point Evaluation – Coding Sheet and Manual.....	34
4.2.2	Content of the Four Coding Blocks	36
4.2.3	Interrater Reliability.....	38
4.3	Data Analysis and Interpretation; Presentation of Results.....	38
4.4	Replication Using Data from Stayner et al. (2007) – First Replication Phase	41
4.4.1	The Overall Effect Size.....	41
4.4.2	Exposure Response Analyses	44
4.4.3	Sensitivity Analyses	47
4.4.4	Publication Bias	47
4.5	Replication Using Data from the Primary Studies – Second Replication Phase....	48
4.5.1	The Overall Effect Size.....	48
4.5.2	Exposure Response Analyses	49
4.5.3	Sensitivity Analyses	50
4.5.4	Publication Bias	51
4.6	A Crossroad Model.....	51
4.6.1	Description of Crossroads.....	51
4.6.2	Finding the Upper and Lower Limits.....	55
4.7	Advanced Analyses	56
4.7.1	Quality of Primary Studies.....	56
4.7.2	Correction of Study Artefacts: An Approach by Hunter and Schmidt	57
5	Results.....	64
5.1	Data Basis.....	64
5.2	Replication Using Data from Stayner et al. (2007) – First Replication Phase	64
5.2.1	The Overall Effect Size.....	65
5.2.2	Exposure Response Analyses	67
5.2.3	Sensitivity Analyses	68
5.2.4	Publication Bias	69
5.3	Replication Using Data from the Primary Studies – Second Replication Phase....	71
5.3.1	The Overall Effect Size.....	71

Index

5.3.2	Exposure Response Analyses	74
5.3.3	Sensitivity Analyses	75
5.3.4	Publication Bias	76
5.4	Results of the Crossroad Model	77
5.5	Advanced Analyses	78
5.5.1	Quality of Primary Studies.....	78
5.5.2	Correction of Study Artefacts.....	79
5.6	Summary	80
6	Discussion.....	82
6.1	Interpretation of the Results	82
6.2	Limitations	85
6.3	Conclusion	88
7	Reference	90
8	Appendix	101

List of Figures and Tables

List of Figures

Figure 1 Possible models due to different homogeneity and systematic effect assumptions (lecture meta analysis class, 2008).....	8
Figure 2 Exemplary meta analysis process in five steps, modelled on Cooper (1982) with several practical examples.	10
Figure 3 Example for a funnel plot. On the left side a study sample without a bias, on the right side a biased sample (lecture meta analysis class, 2008).	13
Figure 4 Theoretical derivations of the CRM and the corresponding literature.....	26
Figure 5 Illustration of a single crossroad and its possible influence of the results of an ETS study.	28
Figure 6 Overview of the coding basis for this thesis. For each block the number of items is reported.....	36
Figure 7 Exemplary 2x2 contingency table as used for the calculation of relative risks or odd ratios. G possibly represents people having a disease vs. people not having the disease. O may stand for a kind of environment exposure present or not present. a to d stands for the (real) distribution of people belonging to one group or another. n are sums of the groups.....	39
Figure 8 Duration exposure response analysis as displayed in Stayner et al. (2007), second figure. Points displayed in this graph are the base for the direct replication in this thesis.	46
Figure 9 Different possible groups of ETS exposure.	52
Figure 10 Display of a number of possible crossroads at evaluation of data stage and data analysis stage.....	54
Figure 11 Forest plot first replication phase.	66
Figure 12 Comparison of the forest plot displayed in Stayner et al. (2007) on the left hand side and a version sorted by the size of the ES on the right hand side.....	66
Figure 13 A plot of the replicated meta regression of the first replication phase. The upper regression line represents a line as it occurs when entering Stayner et al. (2007)	

List of Figures and Tables

regression parameters into the dataset of the first replication phase.	68
Figure 14 Comparison of the original funnel plot of Stayner et al. (2007) on the left and the exact replication on the right hand side.....	70
Figure 15 The forest plot based on the data coded directly from the 22 primary studies.	73
Figure 16 The same forest plot as displayed Figure 15, ordered by the ES magnitude.	73
Figure 17 Replication of the meta regression predicting the OR through the duration of ETS exposure at work, second regression phase.....	75
Figure 18 A comparison of a funnel plot using the inverse variance (left hand side) and one that is using the SE (right hand side), both based the data of the second replication phase.....	76
Figure 19 Forest Plot displaying the good any bad guy analysis as well as the four theories for the Crossroad Model.....	78
Figure 20 Forest plot of all mean ESs. Differences of the replication process and the crossroad model can be compared.....	81

List of Tables

Table 1 The 35 MOOSE criteria for a well documented meta analysis and their appliance to Stayner et al. (2007)	18
Table 2 Key Study Design Features, ORs and 95% CIs for Lung Cancer, second replication phase	72
Table 3 ORs and 95% CI base for the highest intensity analysis in the second replication phase	74
Table 4 Results of the correction of reliability on both the independent and the dependent variable side. On the left side the original parameters are displayed, marked with OR _o and CI _o . On the right side the corrected parameters are marked with OR _c and CI _c	80
Table 5 Comparison of the original study and the two replication phases	81
Table 6 Appendix Content	101

Abstract

The central subject of this thesis is the subjectivity in meta analyses, having an effect on the robustness of results. For this purpose a crossroad model was introduced, crossings thereby representing the subjective decisions a conductor of a meta analysis has to make during the meta analysis process. A meta analysis by Stayner et al. (2007) was replicated by firstly using the data provided in the original study and secondly by collecting the same primary studies and repeating the same analysis steps. The replication produced mostly the same results, but several procedures could not be replicated due to missing information. A sample of possible crossroads was chosen to directly analyse the effect of decisions. A theory free “bad guy” approach was used to show the maximum impact of all crossroads combined (odds ratio= 1.03; confidence intervall: 0.92- 1.15). Four more realistic approaches were chosen to show the effect of decisions in a real world setting (maximum result OR= 1.45 (CI: 1.05; 1.99); minimum result OR= 0.89 (CI: 0.63; 1.27)). The quality of primary studies was added as a moderator, indicating a lower relative risk with increasing study quality. Finally the effects sizes were corrected for reliability of the dependent and independent variable by the artefact approach (Hunter & Schmidt,2004), which lead to an corrected OR= 1.40 (CI: 1.13- 1.73). Limitations and conclusion are considered in the discussion section.

1 Introduction

„It would appear, then, that the major information problem facing psychology is not so much that psychology is producing more information than its total manpower can assimilate, but rather that the individual scientist is being overloaded with scientific information.“ (Garvey & Griffith, 1971, S. 350).

This citation expressed an evolving problem, which was supposed to be solved by a newly introduced summarising method, the *meta analysis*. This method concentrated on quantitative results and offered a more systematic approach, compared to earlier approaches (Bosnjak & Viechtbauer, in press). Correspondingly since 1974 the number of meta analyses published grew exponentially (Schultz, 2004). They were supposed to clarify study questions under dispute by quantitative and objective measures. Formerly, an overview of a research topic was achieved by narrative reviews, which collected some of the studies available on one research topic (primary studies) and came to an overall conclusion, expressed by the according author. This method is susceptible to subjectivity. Often reviews of the same topic came to different conclusions, depending on the author's opinion and the studies the review was based on.

Since the first introduction of the phrase meta analyses by Glass in 1976, they were not only used in psychology research. Other sciences started implementing the method as well, e.g. the medical science. One example is the discussion of harmfulness of passive smoke (Biggerstaff, Tweedie & Mengersen, 1994; Chappell & Gratt, 1996; Le Vois & Layard, 1994, Stayner et al., 2007). Yet, taking a closer look on these publications, it becomes apparent that meta analyses as well can lead to different conclusions. The earlier analyses resulted in a non significant result, whereas the newest meta analysis by Stayner and colleagues resulted in a significantly increased risk of lung cancer. There is a need to disclose the source for these differences, especially when research has a direct impact on legislative provisions. In 2007 a draft law in the Deutsche Bundestag (Lower House of German Parliament) was passed, which aimed at the hazardous side effects of passive

Introduction

smoke: „Schutz vor den Gefahren des Passivrauchens, das nach gesicherten wissenschaftlichen Erkenntnissen für schwere Erkrankungen und Todesfälle ursächlich ist.“ (Deutscher Bundestag, Drucksache 16/5049, 2007, p.1). Roughly translated this means, that the introduction of safety measures because of the danger of passive smoke inhalation is needed, as *assured scientific results* have proven the responsibility of Environmental Tobacco Smoke (ETS) for serious illness and death.

How solid are these “assured scientific results”? Can legislation be passed based on them? Can medical precautions be encouraged by them? Can meta analysis really be the “rock solid” foundation of science (U. Keil, symposium contribution, 23. October 2007), as it is sometimes claimed?

Throughout the hereafter presented thesis it shall be analysed, whether two researchers which were asked to do the same meta analysis on the same topic with the same underlying primary studies will always reach the same conclusion. The influence of subjectivity on meta analysis is examined in a new frame, the crossroad model. Crossroads represent different decisions a conductor of a meta analysis has to make. They outline the impact different positions of points have on the results. The questions arising from this model are addressed to a real world problem, the connection of ETS exposure and lung cancer. A recent meta analysis, conducted by Stayner et al. (2007) was replicated, the crossroad model and advanced meta analytical procedures were applied on the meta analysis to shed more light on the influence of subjectivity on a common research topic. Thereby the main focus was laid on the methods, whereas conclusions about the research topic per se are secondary.

2 Theoretical Background

Some sources date the first meta analysis as far back as 1904, when K.Pearson aggregated some correlation coefficients about a vaccine which was meant to help British soldiers abroad (Hedges, 1987; Shannon, 2008). A more serious attempt was introduced by Beecher in 1955 (Spector & Thomson, 1991; Ziegler et al., 2004), the phrase meta analysis however was not introduced until 1976 by Glass. He stated that „Meta analysis refers to the analysis of analyses. I use it to refer to the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings” (Glass, 1976, p. 3). Independently from Glass, Hunter and Schmidt (1977) developed a similar approach.

Before the introduction of meta analyses the common way to summarize the findings of a specific field of research was the narrative review (Cooper & Hedges, 1994). Traditionally these reviews are based on a selection of literature concerning the research topic at hand. Yet, because of this unsystematic literature research this way of summarizing results is receptive to the influences of the author’s opinion: “Integrative reviewers have not been obligated to apply *any* standard analyses and interpretation techniques in their synthesis process. Most frequently, reviewers interpret data using inexplicit rules of inference” (Cooper 1982, p 297).

Generally, meta analyses have a reputation of being more objective compared to literature reviews. However, this tool has its disadvantages. Many authors have exclaimed problems that might appear, solutions to some of these problems were found. Below an overview of the method meta analysis can be found, as well as a critical discussion.

2.1 Meta Analysis – Development and a Five Step Approach

After a debate about the beneficial effects of psychotherapy, caused by Eysenck in 1952 and resulting in a huge amount of research with partially contradicting results, the need of an instrument to aggregate these finding became apparent. The approach introduced by Smith and Glass (1977) statistically standardized and averaged treatment and control differences for 375 psychotherapy studies (Lipsey &

Theoretical Background

Wilson 2001). Since then, the method “meta analysis” became a growing research field of its own, as a way to replace literature reviews or other forms of summaries like the voting method (Schulze 2004; Hunter & Schmidt 2004). Cooper (1982) proposed a five step model of essential stages to every meta analysis: (1) problem formulation; (2) data collection; (3) evaluation of data points; (4) data analysis and interpretation; and (5) presentation of results. These steps will be dealt with throughout this paper.

2.1.1 Problem Formulation

The first planning step of almost every behavioural science inquiry sets the limits of the topic to be researched. What seems to be an easy task at first becomes more complex, when dealing with research fields where formulations might vary and nominations differ from author to author. Two kinds of definitions, conceptual and operational definitions, have to be established. Conceptual definitions describe the underlying concepts of an analysis; they vary in abstractness and distinguish the analysis at hand. At this point hypotheses or research questions are stated. These definitions are important to deal with as it resolves in part the incommensurability problem (or Apple and Oranges Problem). By clearly defining the level of aggregation misunderstandings concerning the broadness of the results can be avoided. Operational definitions on the other hand describe the relation of abstract constructs with observational variables. Reviewers, opposed to primary analysts, can include additional operationalizations during the process of the review. This usually leads to the inclusion of many different operationalizations in one given review, depending on the reviewer’s choices. This will play an important role later in this thesis (see section 2.3.1). According to Cooper (1982) this is mainly a threat to validity, as some reviews choose to limit their range of operational definitions. This appears either in a problematic externalization of the results or in a further multiplication of concepts, which are not satisfyingly operationalized. Adding more operational definitions only increases the validity, if the additional data is correctly coded and accounted for, e.g. in enough detail. A key roll lies with reporting the underlying definitions without exception, readers of the review should be able to

reproduce the work's foundation.

2.1.2 Data Collection

A literature research needs to be performed after setting the specifications of the meta analysis to be studied (Cooper 1982). Criteria need to be specified, clarifying which primary studies will be included and which excluded. These criteria are based on the operational definitions of the problem formulating stage (Abrami, 1988). To achieve a representative study sample for the study question at hand is the most serious problem of this analysis stage. Rustenbach (2003) recommends a complete inventory count, White (1994) an exhaustive literature research. This research should not only include published articles but dissertations, unpublished scripts and other sources of data as well. All possible sources need to be taken into account, e.g. electronic data bases, literature indices, references of similar studies, specialised libraries and even direct consultation of colleagues, experts and authors. Attention must be paid to interdisciplinary different sources to avoid a biased sample. DeCoster (2004) proposes a "master list" in which all apparently applicable primary studies are listed, exclusions should be executed later.

Problematic for this stage is the fact that publication of an article is more probably, if the results are in line of the common opinion and significant (Rothstein, Sutton & Bornstein, 2005). This produces a systematic bias often referred to as "publication bias". This will be dealt with in more detail in section 2.1.7.

2.1.3 Evaluation of Data Points

By the evaluation of data points, information enclosed in the primary studies of a meta analysis are transformed into quantitative values. Thereby at least three information areas have to be covered (Rustenbach 2003). First of all, data concerning the study question is required, e.g. interventions or basic conditions. Secondly, variables which could bias the results have to be coded, such as quality issues or the used study design. Finally, general study characteristics need to be coded, the year of publication, the authors or the publication media. Which items exactly are to be coded in the end should be guided by theoretical assumptions and

Theoretical Background

the hypotheses stated in the first meta analysis phase (Lipsey, 1994). Though it is recommended to determine the items to be coded in advance, the system should be flexible enough to add additional items if the need emerges during the actual coding phase. Every minor difference between two studies can lead to a knowledge gain concerning the relationships in question.

During the third phase of a meta analysis two documents should be developed: A coding sheet and a coding manual. While the sheet will hold the later coded values of the primary study, the manual describes the variables to be coded in detail. Each coder must have the same understanding of an item, otherwise a bias caused by different coder opinions will be introduced to the analysis (Rustenbach, 2003). The use of several trained and independent coders is recommended to increase the validity of coding. Ideally the coders should be blind to the author, the origin and the hypotheses of the study. When the coding is performed by each coder independently an intercoder reliability can be calculated and the "Effect Sizes" (ES) can be controlled and corrected if needed (DeCoster, 2004). Intercoder reliability should be calculated at the end of the process to show the overall quality of the data, however additional calculations during the process can help to identify problematic items and clarify the problems that cause the different values (Lipsey & Wilson, 2001). One common value to calculate the interrater reliability can be found in Krippendorff's Alpha (Bosnjak & Viechtbauer, in press). This parameter compares the variation between different coders to a random variation. It can be adapted to several different coders and to different scales of measurement and is thereby a very flexible tool to identify the intercoder reliability (Lombard, Snyer-Duch & Bracken, 2004). A high reliability is considered for an alpha between 1.0 and .8, an average for .8 to .6 and a poor reliability for results smaller than .6. Here the items should be newly defined and described (Endrass, Rosegger & Urbaniok, 2007).

Another threat to validity is the incomplete reporting of data. Though this problem could be easily be solved by contacting the original authors of the concerning primary study, such attempts often are not successful. Authors are often reluctant

to answer questions which could threaten their outcomes (Bosnjak & Viechtbauer, in press). Even if the missing data could not have been obtained, there are ways to compensate. See section 2.5.2 and Hunter and Schmidt (2004) for more details.

2.1.4 Data Analysis and Interpretation

The main goal in this step is to synthesize the data into a unified statement about the research problem (Cooper, 1982). Most of the time this is accomplished by performing a weighted aggregation of the data, e.g. the mean ES with accordant CIs.

Some authors state a need for a homogenous study group (Bortz & Döring, 2006), however there are ways to address the problem of homogeneity. Before conducting a meta analysis several questions need to be asked (Schulze, 2004):

- A) Is there a good reason to assume a “one and only” kind of effect underlying the field of study proposed for the meta analysis, or are there any differences between the characteristics of them?
- B) If some differences between the chosen studies have to be assumed, are these differences based on moderating effects and thereby theoretically discoverable and correctable or on unsystematic effects typical for the field in question?
- C) What kind of inference is intended? Is the intention to generalize the results of the meta analysis to all potential studies in the field or to present results based only on the studies included?

As a result of these questions four different models emerge, the so called: fixed effect models, fixed effect models with moderators, random effect models, mixed models or random effects with moderators, respectively (see Figure 1).

		<u>Moderators</u>	
		No	Yes
<u>Random Heterogeneity</u>	No	Fixed-Effects Model	Fixed-Effects with Moderators Model
	Yes	Random-Effects Model	Mixed-Effects Model

Figure 1 Possible models due to different homogeneity and systematic effect assumptions (lecture meta analysis class, 2008).

Here again it is essential to clearly point out the choice of model to the possible audience of a meta analysis, as it has extensional influence on the interpretation of the results. However, often the choice is not made at all, but a procedure is adopted from textbooks of predecessors without asking the questions mentioned above. Here a lot of awareness is still to be risen (Schulze, 2004).

Besides that problem, there is a dispute between different meta analysis branches, where some claim scientific analysis is only possible through random effect models, as significant fixed effect models results might be due to the coincidental choice of studies and not to the overall effect of the entire study universe (Field, 2003; Hunter & Schmidt, 2004). Opposed to this mindset, meta analyses based on fixed effect models are still the most common (Field, 2003; Schulze, 2004).

There are several ways to check quantitative analyses for possible artefacts, labelled with the general term “sensitivity analyses”. Basically the same analysis is repeated with different methods to expose possible effects of used procedures (Cochrane Collaboration). Additionally, this idea can be expanded to content related matters. For example the exclusion of single primary ESs is quite common to discover studies with the biggest impact on the results, thereby discovering possible outliers. In addition, there are different methods to replace missing values in the data set at hand. The emphasis on sensitivity analyses should be bigger depending on the

number of underlying studies. The smaller the study sample, the bigger the influence of a single value and the bigger the misleading influence of a biased method. If on the other hand sensitivity analyses show rather small impact in study results, they can be interpreted as more robust and credible (Bosnjak & Viechtbauer, in press).

2.1.5 Presentation of Results

This often somewhat neglected topic of meta analyses needs to ensure a correct interpretation of the results. In order to assure this, a step by step description of the used procedures is essential. A third person should be able to replicate the very same project easily. This is true for every scientific approach. However for meta analyses some special information has to be accounted for. Especially the foundation of the analysis, the underlying primary studies, need to be described, how they were found and what data was extracted from them (see 2.1.1. and 2.1.2). Additionally some statistical procedures used in meta analysis are still quite new and object of ongoing quality discussions. Here a detailed description of the methods used opens the possibility to retrace the process and enhance the procedures if new techniques become available.

Related to this topic another problem emerges: The interpretation of meta analysis results is somewhat difficult in comparison to most primary studies, since depending on the used model and effect sizes a different mind frame is needed. For meta analytic calculations values without allocation bases are needed, however interpretation of such results are difficult. Gigerenzer and colleagues published some pertinent work about this topic, especially concerning the adequate communication of “relative risks” (RRs)(Gigerenzer et al., 2008). A translation into more vivid measures will be needed if an audience beyond the scientific world is supposed to be reached. Moreover, interpretation of results as they appear in a meta analysis on a personal level, might be misleading. Meta analytic results are based on highly aggregated data, an interpretation on the wrong aggregation level is not possible (Wittmann, 1988). Last but not least, causality of results is in meta analysis as uncertain as in other correlation related procedures. The two last problems

Theoretical Background

mainly concern the moderator analyses (Bosnjak & Viechtbauer, in press).

A general overview of a common meta analytic process can be found in Figure 2.

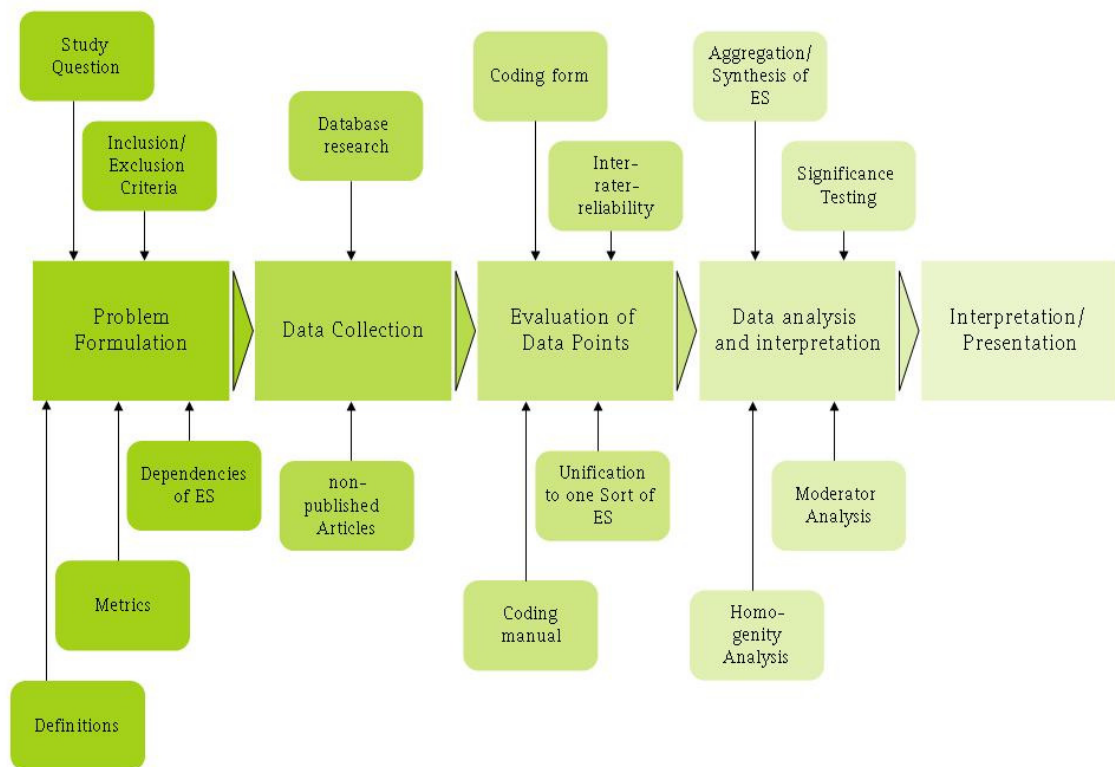


Figure 2 Exemplary meta analysis process in five steps, modelled on Cooper (1982) with several practical examples.

2.1.6 Different Meta Analyses Approaches

It is somewhat misleading to talk about *the* meta analysis approach. In addition to the general approach described above, several families or approaches of meta analyses have been introduced: the “p” family, which can be added to the methods of vote counting, the “Glassian” family, the “Sampling Error” family, and the “Psychometric Hunter and Schmidt” family (Schulze, 2004). Of course this is only a rough selection of authors and families, due to the intensive process of the method “meta analysis” since 1971 there are plenty variation and additions to these approaches.

For this thesis the sampling error family, often related to Hedges and Olkin (1985)

and the Hunter and Schmidt approach (2004) are most important. Hedges and Olkin tried to solve the problem caused by the sampling error by introducing the concept of homogeneity tests. If heterogeneity is found in a study sample, study parameters might be based on different population parameters, an interpretation of the results would then be impossible (Fricke & Treinies, 1985). Hedges (1982) proposed a χ^2 -test to check for homogeneity of the study sample at hand. This test compares the given variation of ESs between studies to the expected variation due to the sampling error. A significant test would mean that e.g. additional moderators are needed to clarify the variation not explained (Hunter & Schmidt, 2004).

The Psychometric approach (Hunter, Schmidt & Jackson, 1982) focuses on correlation based ESs. Hunter and colleagues propose that unsystematic and systematic bias like sampling error, or a small reliability of the measured variables influence the results of a meta analysis. Depending on the size of these errors a correction can be applied to the analysis. This approach is described in more detail in chapter 2.5.2 and 4.7.2.

For a more detailed overview of techniques and methods of different meta analyses, see Cooper, Hedges and Valentine (2008).

2.1.7 Common Critique Concerning Meta Analyses

As with any other method, there are problems and biases which have to be addressed when dealing with meta analyses. Most, but not all of them, are related to the choice of the studies added to the study sample. The following paragraph will introduce some of these issues.

Incommensurability: This problem concerns the assimilation of different kinds of studies into a single output variable. Critics emphasize the impossibility to add up a larger amount of scientific studies, as they derivate too much to be able to extract one single information from them. Even though studies might appear to cover the same topic, underlying concepts and constructs of studies often vary significantly due to scientific freedom and different operationalizations. Thus “apples and oranges” would be added together. Yet, this argument is debilitated easily by

Theoretical Background

adapting the view to fruits instead of apples and oranges or, to put it differently, by switching to the right level of aggregation (Wittmann, 1988). To avoid information loss, the properties of apples and oranges can be added as moderators. Here the evaluation of data points (see 2.1.3) is especially important, since special features of the studies need to be identified and coded in order to draw conclusions from different aggregation levels. The result is a clear picture of how apples and oranges differ, showing how they are similar and different (Bosnjak & Viechtbauer, in press; Franke, 2001).

Publication bias: This frequently raised objection addresses a bias caused by certain publication practices. Many scientific journals have explicit or implicit restrictions concerning which articles will be published and which not. To have a paper published when the results are not as expected or insignificant is less probable, as compared to significant and conform results. Other sources refer to this problem as “file-drawer” problem, expressing that these studies never leave the researcher’s office (Rosenthal, 1979).

Several methods are available to either discover a present publication bias or to express its magnitude. Probably best known are the calculation of the “Fail-Safe N” and the “funnel plot” (Bosnjak & Viechtbauer, in press). The first is estimating the number of insignificant studies needed to push the results of a meta analysis below a certain magnitude (Rosenthal, 1979). This approach and newer versions of it are developed in Pahl (2009). The second is a more graphical solution where the effect sizes are plotted against different kinds of precision measurements, which are in turn related to the sample size of the study. As the spread of the effect sizes should increase in accordance to the precision of measurement, a typical funnel should be discoverable around the mean ES if no bias is to be found in the data set (see Figure 3, left side). On the other hand, if there is a bias a typical gap should appear in the low ES part of the plot (see Figure 3, right side) (Light & Pillemer, 1984). For a more detailed description see the method section (4.4.4) and Pahl (2009). All these methods have a somewhat subjective interpretation in common: Which fail-safe N is high enough or which funnel plot reveals a bias remains in the eye of the

beholder.

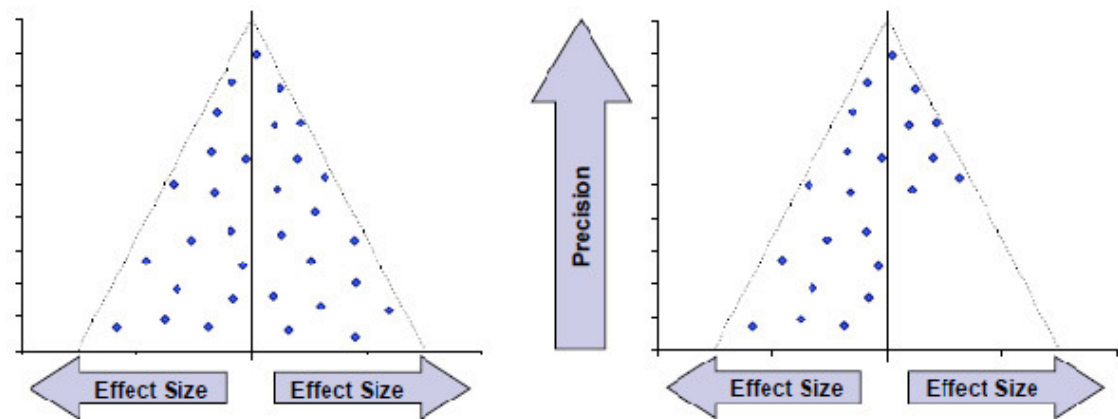


Figure 3 Example for a funnel plot. On the left side a study sample without a bias, on the right side a biased sample (lecture meta analysis class, 2008).

Quality issues of primary studies: Another problem may be uncovered in case studies of either different or overall poor quality are aggregated, thereby threatening the internal validity (garbage in – garbage out). A way to avoid this problem might be to set a standard for studies, in terms of sufficient internal validity and power, to be included for the meta analysis (Bortz & Döring, 2006) (see section 2.1.2). However, this comes with a limitation for external validity and needs to be balanced against each other. Additionally meta analysis should rather be a possibility to include low power studies into the equation (according to their study size) instead of excluding them.

Another way to deal with this problem can be found in Wittmann and Matt (1986), where quality issues were operationalized guided by the four validities as described by Cook and Campbell (1979). Here quality issues showed to have a major impact on the results of a meta analysis. In contrast Hunter & Schmidt (2004) incorporated quality as a weighting factor to exclude the influence of quality in the analysis step of a meta analysis. This approach will be dealt with more closely in chapter 4.7.1.

Adjustments: One more problem can be found in the more or less common practise to adjust results of primary studies for confounding effects like the influence of gender or age effects (Bosnjak & Viechtbauer, in press). Depending on the topic many other adjustment variables are imaginable, besides the possibility to not

Theoretical Background

adjust at all. Thus finding a homogenous study sample where all studies use the same adjustments is close to impossible. Again, there are several approaches to deal with this issue. One is to control for possible differences by adding adjustment vs. no adjustment as a moderator variable, another is to transform the different ESs into one homogenous form (Peterson & Brown, 2005). At least it should be documented closely which adjustments are present in the data set at hand and how this problem was addressed so far, but as for a generally advisable approach, this one has still to be found. How the problem is addressed in this work can be found in the chapters 2.4 and 4.6.1.

Dependent ESs: One of the major assumptions needed for a meta analysis is the assumption of independence between the according ES. However, often primary studies will report more than one ES. Besides the advantage of having more data to add to the meta analysis, this violates the assumption of independence of a standard meta analysis (DeCoster, 2004; Hunter & Schmidt, 2004). Additionally, several publications might be created from the same data set or the same “research environment”, consequently producing dependence much harder to discover as compared to the multiple ESs of a single study (Bosnjak & Viechtbauer, 2009). The easiest and most often used form to deal with this issue is to simply ignore it. One indicator for such a meta analysis is an ES number higher than the number of primary studies. There are other ways to compensate for this problem, though: Aggregating ESs for each study if several are reported or (more or less) randomly choosing one of the reported ESs. However closing the problem on dependency disregards a lot of information. It would be recommendable to model the influence of dependence in the given dataset to both ensure the assumption of independence and use the biggest amount of information possible for the analysis. Sadly this procedure is not yet found in present meta analyses frequently (DeCoster, 2004; Bosnjak & Viechtbauer, 2009).

2.2 Stayner (2007) - Lung Cancer Risk and Workplace Exposure to ETS

Many approaches for a statistical analysis of the method meta analysis use so called “Monte Carlo” data (e.g. Gilpin, 2008). This is a reasonable approach, as the

Theoretical Background

supposed results of a possible meta analysis are known and the diversion from these supposed results can be measured as a way to define the quality of meta analyses as a quantitative method (Gilpin, 2008; Koetse et al., 2007; Sánchez-Meca & Marín-Martínez, 1997). In this paper a published meta analysis was chosen to emphasise the real world effect of the proposed study questions (see chapter 3). In addition an intensively discussed topic, the exposure to ETS was chosen, to show how even in highly researched areas the variations of possible scientific decisions are of a significant effect. Decisions and their subjective nature may change depending on the study topic at hand and would never emerge in a purely theoretical framework. Thus a “real world” problem might shed more light on subjectivity.

For these reasons, a meta analysis conducted by Stayner, Bena, Sasco, Smith, Steenland, Kreuzer, and Straif (2007) is used in this paper as an example for a meta analysis. All further data will be derived more or less closely from the content of this study. This paper is meant to propose new methodological approaches to meta analyses about ETS. Thereby some deeper insights into the study field might have been missed; emphasis was laid on methodological research aspects of ETS in the work environment.

2.2.1 The Field of Research of “Environmental Tobacco Smoke (ETS)”

The consequences of smoking cigarettes and other tobacco products on health are commonly known by now. Scientific prove of the harmful effect has been build up since the 1950s (Doll & Hill, 1950) and has a stable foundation today. Taking this evidence into account it is plausible to assume that smoke emitted directly from a burning cigarette and inhaled by a present non-smoker will contain the same poisonous particles and will have therefore the same harmful effect (Hackshaw, 1998). For this reason numerous studies have been conducted analysing the association between ETS and the risk of lung cancer for non-smokers, e.g. at home or in the work environment. However, here a different picture arose, results were not as clear or strong as expected, given the strong evidence in direct smoking studies. Some studies found a quite raised risk of lung cancer when cases were

Theoretical Background

exposed to ETS (e.g. Reynolds et al., 1996), others could not reject the null hypothesis (e.g. Wang, Zhou & Shi, 1996). Here meta analyses were introduced to clarify the view. Nevertheless, again differentiating results were found (e.g. Hackshaw, Law & Wald, 1997; Le Vois & Layard, 1994).

2.2.2 Stayner (2007) Lung Cancer Risk and Workplace Exposure to Environmental Tobacco Smoke – a Short Review

Stayner et al. (2007) see the reason for a new meta analysis with the topic of ETS and lung cancer in the inclusion of some new evidence and in the addition of some sophisticated methods. Consequently this was rather meant to be an expansion of the present evidence, not a completely new approach. The data basis for the meta analysis from Stayner and colleagues was found in former meta analyses and an additional literature research in Medline and Embase conducted on the 1st of January in 2003. Overall a number of 22 studies were identified as feasible for their meta analysis, resulting in 25 ESs. As a rule they proposed to use adjusted values whenever possible, however this rule could not always be maintained, a number of unadjusted values were used. RR, “confident intervals” (CIs), and important study attributes were coded. On this data base numerous analysis methods were used, described in more detail in section 4.4.

2.2.3 Main Results of Stayner et al. (2007)

In this section an overview of the major results will be given. Methods adopted by Stayner et al. (2007) can be found in greater detail in the replication section of this thesis.

The described relative risk based on the 22 primary studies was 1.24 (CI: 1.18; 1.29) with an underlying fixed effect model and 1.24 (CI: 1.17; 1.31) with a mixed effect model. α was 5% in both cases. No signs for heterogeneity were found via random-effects model ($p=.08$) or DerSimonian-Laird test ($p= .49$).

As moderators a conglomerate of so called “key study variables” were tested, among which only the variable “the study controlled for exposures to other occupational carcinogens” was found to be significant, indicating a higher relative

Theoretical Background

risk (RR= 1.59) for those studies which controlled for this effect as compared to studies without this correction (RR= 1.14).

Furthermore two exposure – response moderators were analyzed: First the intensity of ETS exposure, second the duration of the exposure. For the first case there was a limitation of seven studies which reported the needed information, for the second case six studies were available. Aggregating seven values of the highest intensity lead to the overall result of a RR of 2.01 (CI: 1.55; 2.60) with the fixed effects model and a RR of 2.01 (CI: 1.33; 2.60) with the random effects model. Here again no evidence for heterogeneity was found. For the relation of duration and relative risk a meta regression was conducted. The resulting regression equation with a slope of .011 and a standard deviation of .0025 leads to a prediction of a relative risk of 1.63 (CI: 1.45; 1.82). The prediction was reported to be highly significant ($p < .001$). A graphical implementation can be found in Figure 8.

Stayner et al. (2007) used a number of sensitivity analyses to identify outliers in their basic data set. The influence of single studies on the overall result and on the intensity – relative risk moderator analysis was tested, in addition to the exclusion of a number of studies which were introduced to Stayner et al. (2007) but missing in Wells (1998). None of these tests showed a significant influence on the according result.

Publication bias was examined through a funnel plot (see Figure 14), in which the relative risks of the studies were plotted against the multiplicative inverse variance of the logarithmized relative risks. At the discretion of Stayner et al. (2007) there is no evidence of a publication bias to be seen in this plot.

2.2.4 Flaws in Stayner et al. (2007)

On closer inspection of Stayner et al. (2007) some limitations of the methods and results became obvious. In the following several of these major problems will be described. The subsequent replication will show the size of the effect these flaws have. It is important to emphasize that only methodological issues are accounted for; no content related problems were addressed. However, as the incomplete or

Theoretical Background

wrong use of methods will have a major impact on the later interpretation of the content it is of the utmost importance to report the used procedures and to stick to some rules. A conglomerate of these rules is found in Stroup et al. (2000) „Meta-analysis Of Observational Studies in Epidemiology” or the “MOOSE” group in short. Here the ideal meta analysis is divided into six parts, each having several subsections. When the resulting 35 points are applied to the work of Stayner et al. (2007), only 8 were fulfilled and 11 were met to some extend. 16 principles were not covered at all. An overview of the checklist provided in Stroup et al. (2000) and the points Stayner et al. (2007) met is given in Table 1.

Table 1 The 35 MOOSE criteria for a well documented meta analysis and their appliance to Stayner et al. (2007)

Six major elements of meta analyses	Content of these six elements	Amount of description in Stayner et al.
Reporting of background	Problem definition	Partially
	Hypothesis statement	None
	Description of study outcome(s)	Partially
	Type of exposure or intervention used	None
	Type of study designs used	None
	Study population	None
Reporting of search strategy	Qualifications of searchers	None
	Search strategy, time period included in the synthesis; keywords	None
	Effort to include all available studies, including contact with authors	Partially
	Databases and registries searched	Complete
	Search software used, name and version, including special features used	None
	Use of hand searching	Complete
	List of citations located and those excluded, including justification	Partially
	Method of addressing articles published in languages other than English	None
	Method of handling abstracts and unpublished studies	None
	Description of any contact with authors	Partially
Reporting of methods	Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested	None
	Rationale for the selection and coding of data	None
	Documentation of how data were classified and coded	None
	Assessment of confounding	Partially
	Assessment of study quality, including blinding of quality assessors; stratification or regression on possible predictors of study results	Partially
	Assessment of heterogeneity	Complete
	Description of statistical methods in sufficient detail to be replicated	Partially
	Provision of appropriate tables and graphics	Complete
Reporting of results	Graphic summarizing individual study estimates and overall estimate	Complete
	Table giving descriptive information for each study included	Complete
	Results of sensitivity testing	Partially
	Indication of statistical uncertainty of findings	Complete
Reporting of discussion	Quantitative assessment of bias	Partially
	Justification for exclusion	None
	Assessment of quality of included studies	None
Reporting of conclusions	Consideration of alternative explanations for observed results	Partially
	Generalization of the conclusions	Complete
	Guidelines for future research	None
	Disclosure of funding source	None

The first block described in MOOSE covered which background information should be reported. First of all a definition is needed. In the original publication it is

Theoretical Background

described that several earlier studies concerning ETS and its impact have been conducted but no details of any kind are given. Therefore a close definition is missing. Additionally no hypotheses are reported in Stayner et al. (2007). Some considered precursor studies are described, hence this point is covered to some extent, but no numbers are reported. There are no descriptions of exposure or interventions used to be found in the introduction. Finally neither the study design nor the study population is portrayed.

The second block in Stroup et al. (2000) concentrated on the description of search strategies a meta analysis data was based on. The request to describe the background on the literature search team was not answered in Stayner et al. (2007). No search terms or search strategies are named, but the date of the literature research. The effort to include all available studies of the research field seems to be fulfilled, though very little is reported about this topic. The data bases and registries which were used are mentioned (MedLine Embase). Besides this only the use of reference lists of former meta analysis about passive smoke is mentioned, which can be considered as a hand search. No search programs and their properties are listed. There was no list of excluded studies or a number of excluded studies, only an overview was given of why some studies were excluded. No procedures to include studies of a language other than English are described. Nothing is said about the enclosure of abstracts or unpublished articles. Contact to other authors is described for one case, to solve the problem of shared cases and controls between Boffetta et al. (1998) and Kreuzer et al. (2000). Sadly here the bibliographical reference was misprinted, however in the first table there is a hint which one is the correct Boffetta study.

The third block presented by the MOOSE group regards the methods of a meta analysis. The first three points, the description of relevance or appropriateness of studies assembled for assessing the hypothesis, the rationale for the selection and coding of data and the documentation of how data were classified and coded are all not reported in Stayner et al. (2007). Assessment of confounding took place to some account, as several adjustments of primary studies were coded as key study design

Theoretical Background

features and included into later analyses. However the list was far from complete (see Table 2). Study quality was not mentioned in any constructive way. Heterogeneity was analysed by the likelihood ratio test the test of DerSimonian and Laird (1986) and was thereby dealt with, just like the usage of appropriate graphs and tables. The requirement to describe methods and procedures to an extent that a complete replication is made possible was only met partially, as will be described in chapter 4.4.

The fourth block of Stroup et al. (2000) deals with the presentation of results. Graphical conversion of analyses were displayed, a table with descriptive data is available. Statistical certainty was reported by p-values and CIs. Only the sensitive analysis could have been described in more detail. The studies which had the biggest impact in the leave-one-out analyses were not reported for example.

The fifth block sets rules about the discussion section of meta analyses. Problems and limitations of Stayner et al. (2007) were touched at least partially, they e.g. talk about wrong classification in the 2x2 table for cases and controls. Publication bias was covered without reporting the corresponding results. The justification of excluded primary studies or the assessment of primary study quality is completely missing.

The final and sixth block described by the MOOSE group is about reporting of conclusions. First here alternative solutions for the gained results should be described. This was done to some extent in Stayner et al. (2007) by mentioning several limitations. The level of generalization of results is stated quite clearly: "The findings from this meta-analysis in conjunction with the findings from ETS studies of nonsmoking spouses provide compelling evidence that exposure to ETS in the workplace is a significant risk factor for lung cancer" (Stayner et al. (2007), p. 550). Finally no guidelines for future research are given and the source of funding is not disclosed.

Closing the matter on the MOOSE criteria other method issues were found which are not covered by Stroup et al. 2000. There was no general decision for either the fixed or the random effects model. This fits more to an explorative study than to a

Theoretical Background

confirmatory like the study at hand. Here the level of generalization should be set in advance (Bosnjak & Viechtbauer, in press). The same seemed to apply to the moderator analyses, reading the original article the impression comes up that moderator analyses were conducted fitting to the data set, not according to theoretical assumptions arranged in advance. In addition no procedure for weighting the ESs is described. The weighting procedure which was used in the replication phase (see 4.4) is only based on the assumption that the most common method for weights was used (inverse variance).

Another problem is the inclusion of “never smokers”, which is proposed by Stayner et al. (2007). Yet, the definition of never smokers varies significantly depending on the primary study at hand. There are definitions of “less than 100 cigarettes in life time” up to “less than 400”. Other studies use time period definition, excluding cases who have smoked more than six month one cigarette each day. The biggest percentage of primary studies did not report a definition at all. Concerning the definition of the study sample the inclusion of persons exposed at other places than at work is problematic. These persons will bias the results as they will be exposed to more smoke than the heading of “ETS at work” does imply. This matter is dealt with in more detail later in this thesis (see 4.6.1).

Other incomplete descriptions of procedures rendered the exact replication process impossible. For the sensitivity analyses it is not reported whether studies were excluded or single ESs. Both approaches would be reasonable. Then, as mentioned above, two studies share a part of the cases and controls are in the study sample of the original meta analysis. Though the approach to exclude the double persons is advisably, it is not reported how this effected either the primary study distribution or the final result of the meta analysis. The number of removed persons from Kreuzer et al. (2000) could not be reconstructed and does thereby influence most of the following calculations of this thesis.

Besides these fundamental problems a calculation error was discovered as well. The CIs Stayner et al. (2007) presented around their predicted 45 year exposure period was probably miscalculated. Using equation 19 (introduced beneath) under the

assumption of $\alpha = .05$ and a one tailed test the result should be

$$CI = \exp(45 * (b_p^* \pm t_{krit,df} * s_p^*)) = \exp(45 * (.011 \pm 2.12 * .0025))$$

when using the original data. However this results in limits of $CI^- = 1.2920$ and $CI^+ = 2.0824$ instead of the presented $CI^- = 1.45$ and $CI^+ = 1.82$. These values represent an α level of 68% and probably appeared because it was forgotten to include the critical t-value into the equation above.

Two further mistakes were discovered. First in the subgroup calculation of the highest intensity Stayner et al. (2007) referred to Kabat et al. (1984) although it was data from Kabat et al. (1995) that was included into this analysis. The second was found in table one of the original study. Here the exposure period stated for cases of Kabat et al. (1984) is different depending on the gender. There is no evidence of different exposure periods between men and women in the primary study. Admittedly these kinds of mistakes can be seen as typing error which will be present in any published work.

Regarding the primary studies a flaw appeared several times: The confusion of Odds Ratios (ORs) and RRs, referring to them as the same parameter. However, as will be described in chapter 4.3, the equations of these two ESs forms differ significantly and will influence the results of a meta analysis when one method is chosen above another (Bosnjak & Viechtbauer, in press).

2.3 Subjectivity of Meta Analyses and Implementation of the Crossroad Model

After discussion of meta analyses as a general approach (see chapter 2.1), to many authors it seems clear, that this method is an objective, rigid and somewhat foolproof way to solve questions in diverse study areas (e.g. U. Keil (Symposium contribution, 23rd of October 2007); Morris, 1994; Wanous et al., 1989).

However, there is a rather big fraction in the scientific discussion, which emphasized problems of meta analyses (see section 2.1.7) and some even propose to stop using the method, as it has too many flaws (Eysenck, 1978). Why replications of meta analyses or independently conducted meta analyses about the same topic

lead to different results indeed remains to be explained (Copas, 1999; Wanous et al., 1989).

This thesis is not yet another attempt to decide for one or the other option, discussions and results of the pros and cons of meta analyses can be found throughout the literature (e.g. Eysenck, 1978; Howard et al, 2000; Mundy & Stein, 2008). Instead, a different approach compared to these extreme positions is to use this discussion to the advantage of the general method meta analysis. Within the range and variations of different meta analyses a lot of information can be found, especially if the source of the variation can be discovered. One way to deal with unknown variations of results is the moderator analysis, another way is the heterogeneity approach (Lipsey & Wilson, 2001). Since these approaches depend on the topic in question, as every topic will yield its own moderators and its own amount of heterogeneity, here another source for variations is dealt with: The unavoidable subjectivity influencing the results of any meta analysis.

2.3.1 Meta Analysis and its Subjectivity

“... we cannot help but construct the real, even when it pleases us to think we are doing no more than perceiving it.” (André Green 1986/ 2005, p. 290)

Subjectivity in research is well known and discussed in the literature (e.g. Bradbury-Jones, 2007; Choi, 2006; Peshkin, 1988). In the end, most authors conclude, that the general aim should be to reach as much objectivity as possible, even though it can never be reached to a full extend. A clear definition of subjectivity and its meaning is still object of an ongoing discussion in philosophical cycles. However, for the purpose of this work a definition provided by Peshkin (1988) is quite useful. He quotes “subjectivity as the quality of an investigator that affects the results of observational investigation” (Peshkin 1988, p 17).

Concerning the objectivity and the goal to reach it, there are research methods which are supposed to be more or less objective. A famous example was the introduction of double blind research designs, resulting in a major increase of objectivity in experimental research. Meta analyses were meant to gain the same

benefit for narrative reviews.

In general, quantitative approaches to study questions are often believed to be more objective compared to qualitative approaches. Suggestions about the qualitative aspects to subjectivity can be found in Peshkin (1988) or, for a more health related view, Bradbury-Jones (2007). Here a need to "... systematically identify their [the researcher's] subjectivity throughout the course of their research." (Peshkin, 1988, p. 17) is stated.

So far the main topic about meta analyses problems concerns the selection of the right primary studies to be included into the analysis. This selection problem leads to several different dilemma discussed in 2.1.7. However, looking closely at the discussion in the literature, more sources for subjectivity are mentioned: Different method possibilities, (implicit) subjectivity on a personal level, active influence by the researcher, the source of funding or the institution the research takes place at. An additional source is more related to the concerning topic of research. Often definitions of apparently clear phrases vary between different research groups. For an example see 4.6.1.

Different method possibilities are in the case of the method meta analysis e.g. the different meta analysis schools which have been developed (see chapter 2.1 or Schulze (2004) for a more thorough overview), most of them similar in some ideas how to deal with the problems meta analyses have, but then again quite different in used equations and underlying theoretical constructs. Fixed effects models can be chosen over random effects models, different equations for the estimation of ESs can be used (e.g. Mengersen et al., 1995). Which moderators to use and which to leave often is left to the author of a meta analysis to decide, which will have a huge impact on the final result.

Differences in defining constructs and operationalizations start in the first step of a meta analysis, the problem formulation stage. However, these differences can affect all the other stages as well. Inclusion and exclusion criteria for primary studies need to be defined, the level of generalization of the meta analysis needs to be constituted. Are there only non-smokers to be included? Are there only people

Theoretical Background

exposed to ETS at work to be included? How are cases dealt with, if they were exposed to smoke at home *and* at work? These sometimes not reported decisions might have a major impact on the results of a meta analysis. Already on the level of primary studies different definitions can easily be found (see Koo, Ho & Saw, 1984), however often the definition must be discovered through close inspection (see 4.6.1).

Most difficult to discover are implicit decisions a publisher is not aware of himself. Again this is true for primary studies as well as for meta analyses. Even when for example the quality of primary studies is rated and the process is made as explicit as possible, there still will be some implicit influence in the judgement (Wortman, 1994). Another attribute of these implicit decisions is that they can easily be change on the way. That way in a meta analysis half the primary studies could be chosen by the one set of implicit rules and the other half by another set of rules.

It is important to emphasize that the source of this subjectivity is not lacking ability or the correctness of certain methods compared to others. All these sources of subjectivity have their problems and gains. Often the cause for differences is simply the area, in which the education was build. A scientist with a medical background would more likely use methods related to relative risks, as a psychologist will use correlation related methods. A discussion of the many different approaches which can be found for conducting a meta analysis would go beyond the scope of this thesis, a more detailed description can be found in Lipsey & Wilson (2001) or Schulze (2004).

Concerning more deliberate means to influence outcomes of analysis like the source of funding or personal beliefs of a certain scientist, as e.g. described by Barnes and Bero (1998), we try to focus on less deliberate variables here.

Transferring the subjectivity problem to meta analyses, DeCoster (2004) states that meta analyses are not as much an objective tool but one which provides a “shared subjectivity”, as all steps used to receive the meta analysis results should be reported closely. The main problem is the lack of awareness for the subjectivity problem or the incomplete reporting of the steps taken. As mentioned in chapter

Theoretical Background

2.1.5 a full replication of the concerning meta analysis should be made possible, however in the exemplary meta analysis for this thesis there were difficulties with this demand (see 2.2.4).

Putting all these matters together a model can be derived, to be called “Crossroad Model” (CRM) in this thesis (see Figure 4).

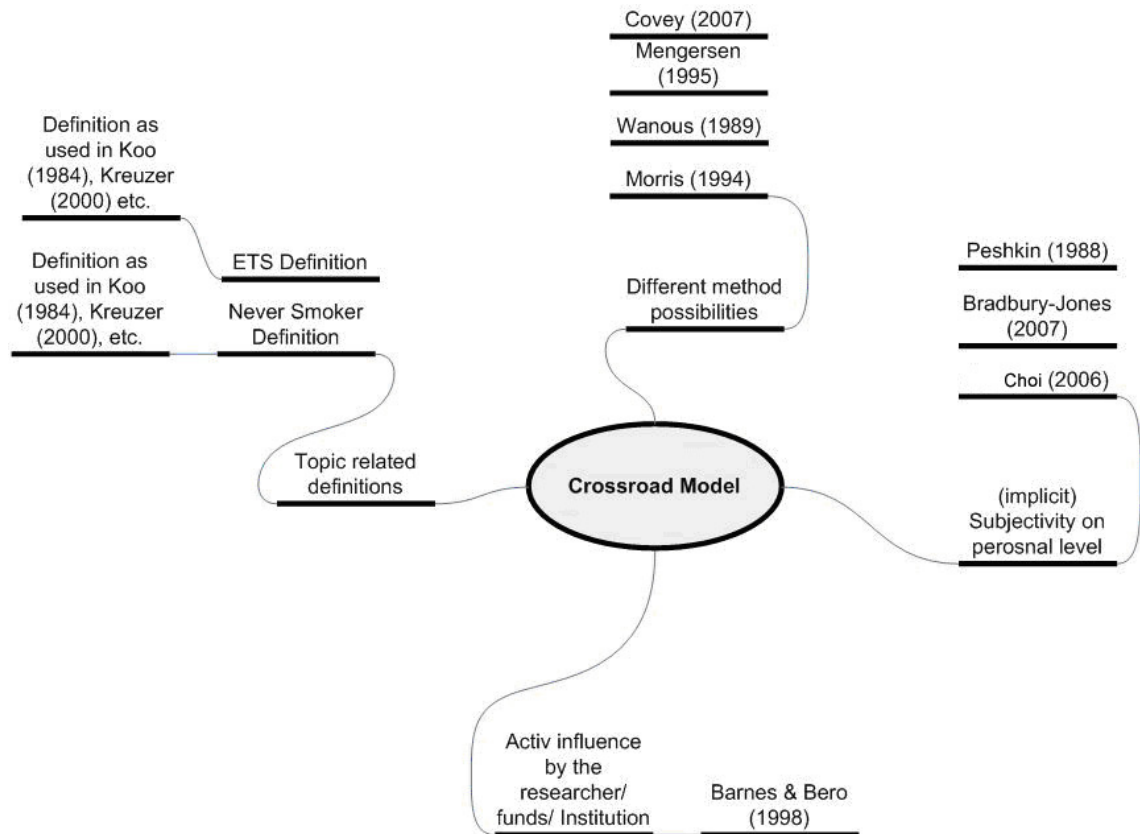


Figure 4 Theoretical derivations of the CRM and the corresponding literature.

2.4 A Crossroad Model

The following scenario shall be the origin for all the following analyses: Imagine two scientists to examine one and the same research question. Both decide to conduct a meta analysis, as there is discrepancy in the topic and apparently enough literature and primary studies available. Now imagine both scientists include accidentally the very same primary studies into their meta analysis. Should not the result of the analysis be exactly the same? This question and similar ones are to be analysed in this thesis.

Theoretical Background

In the following chapters it will be tried to be demonstrated, whether the subjectivity named above does influence the results of a meta analysis and if it does, to what extend. Based on possible subjective decisions a CRM was designed. With regard to its content it is supposed to show the different crossings a research topic and its methods provide to influence the results. Each crossroad represents a decision with several correct options. Leaving from each crossing is at least one way, one different method, one different definition, which will possibly influence all of the following analysis. In addition it will be tried to show which way to take to arrive at a certain conclusion, whichever this conclusion might be. Again here it should be emphasized, that there are no wrong ways in this CRM (of course in reality there are wrong ways like the application of inappropriate methods, however we will not deal with them this work), all ways are in principle correct.

Figure 5 shows an exemplary effect of a subjective decision. On the left hand side the normal five step plan of a meta analysis results into a crossroad. Depending on the subjective decision (displayed here as a switch) the results of the study might vary between a recommendation of a common ban on smoking or an impression of harmlessness concerning ETS.

Even though these crossroads can appear at each step of a meta analysis, in this thesis we will concentrate on the evaluation of data points (see 2.1).

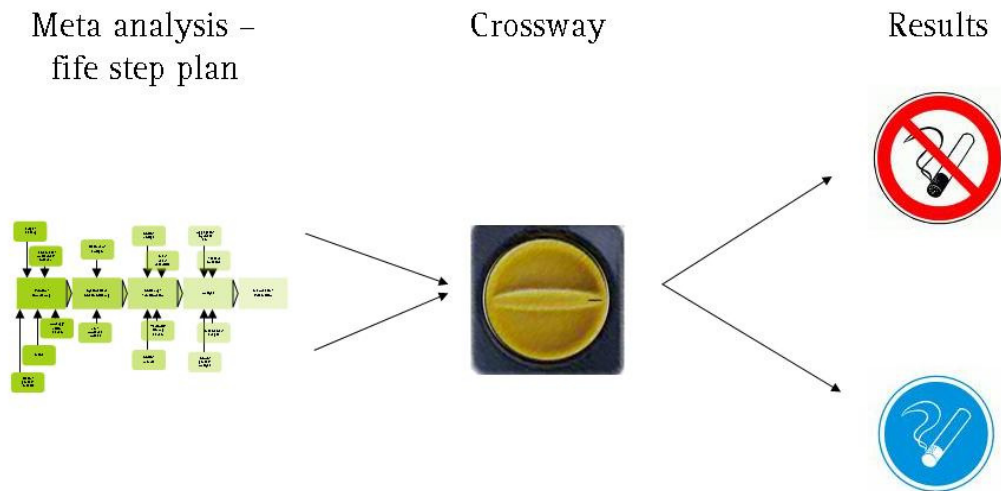


Figure 5 Illustration of a single crossroad and its possible influence of the results of an ETS study.

Besides considerations concerning the CRM further analysis methods will be presented. Without claiming to have exploited the present possibilities completely, it was tried to advance the analysis of the relative risk and ETS at the workplace through additional methods.

2.4.1 Finding the Upper and Lower Limits

Assuming the decisions at crossroads do have an effect on the results of a meta analysis, it is plausible that there is a certain position of points which will lead to one or another extreme value. One might argue that a tobacco lobbyist or an anti smoking activist might come to completely different conclusions, although using the same primary studies and without making methodical mistakes during the meta analysis process. In an attempt to define the size of the effect of the CRM the author and Pahl (2009) tried to find these extreme positions in the present data set and the decisions which lead the way to them.

Studies like Mengersen (1995) have shown, that the choice of appropriate studies for a meta analysis has a big impact on the results of the concerning meta analysis, which could serve as another crossroad. However, in the present study a replication of a former analysis, Stayner et al. (2007), was done in order to shift a bigger focus to other variables influencing results besides this major effect. By using the exact

data of an other meta analysis the range resulting out of one and the same data set can be studied more thoroughly.

2.4.2 About Missing Data

„If there are any ways in which data can be missing, they will be.“ (Cohen et al., 1983, p. 275). This is true for meta analyses as well. Here missing information appears as data which is required by the author of a meta analysis but not reported by a primary study. These amount of this missing data will vary between studies, some will report data quite thoroughly, others rather poorly. This problem is caused in part by the lack of standardisation of the publication processes. Though advice for clear publication can be found in “Strengthening the Reporting of Observational Studies in Epidemiology” (STROBE) (Vandenbroucke et al., 2007) or the MOOSE group (Stroup et al., 2008), these are hardly ever binding. This results in differences between personal styles, editor guidelines or other factors influencing the publication practice (Rustenbach, 2003). Excluding studies with missing information or the use of methods to impute the missing data to the data set are possible ways to face this problem, however it remains to be seen whether any of these methods influence the results artificially.

Introducing the crossroad model causes another source of missing data to emerge. It can happen that a study fits the analysis needs of a meta analysis but is lacking data for a certain decision reached at a crossroad. For some decision paths all studies might be available but for some crossroad combinations the number of studies at hand might differ. Imagine the use of adjusted vs. unadjusted values as one crossing of a meta analysis. All studies might report unadjusted values leaving room for all the other crossroads following this decision. However, only half of the available primary studies report adjusted data, here only 50% of the data will be accessible for analysis.

2.5 Advanced Analyses

The replication of Stayner et al. (2007) is the main method to analyze the CRM and its influence. However, taking a close look at the original meta analysis, it seems

some analysis methods commonly related to meta analyses could have been used to shed more light on the research topic in addition to the analyses Stayner et al. (2007) used. An individual sample of methods is presented in this thesis and in Pahl (2009).

By the way, the choice of the advanced methods can be seen as yet another crossroad, different authors educated at different universities might have chosen different methods. Sadly, time restrictions made it impossible to employ all adequate meta analysis methods to the present data set. For a detailed methods description of the advanced analyses used in this thesis see chapter 4.7.

2.5.1 Quality of Primary Studies

One major point to account for is the quality of primary studies. “Garbage in – garbage out”, an analogy adopted from the computer science and used by Eysenck (1978) warns to not disregard quality issues of primary studies, as all results of a meta analysis will not be meaningful without high quality data.

Having this as a basis the hard part is to define what “quality” really means in terms of good primary studies. In this thesis, two methods were used: One drawn from the STROBE quality recommendations (Vandenbroucke et al., 2007), another was triggered by the threads to validity as introduced by Cook and Campbell (1979). The first quality basis is a mere accumulation or recommendable publication methods or needed information for the reader of scientific, epidemiologic articles. A checklist of 22 items was designed to guide authors in their attempt to publish their data and results in a comprehensible, high quality way. These recommendations were translated into a quality rating (see appendix D and Pahl 2009).

The second method used “Statistical Conclusion Validity”, “Internal Validity”, “Construct Validity of Putative Causes and Effects” and “External Validity” as a guide to design items to control for these validity issues. High ratings on all four kinds of validity would result in a high quality value. This method was used to a high degree in the meta analysis of Wittmann and Matt (1986). For the adaption for this thesis see section 4.7.1.

2.5.2 Correction of Study Artefacts: An Approach by Hunter and Schmidt

As described earlier Hunter and Schmidt developed independently from Glass a meta analysis approach at the same time (Schultze, 2004). Besides many similar assumptions Hunter and Schmidt had a supplementary accessory in their theoretical framework: The correction of artefacts. The base assumption for artefacts is that there are no perfect studies. Every study will be influenced by artefacts to some extend. Thus the analysis of relations between studies is more complicated than just measuring it. In their work Hunter and Schmidt present eleven sources of artefacts influencing the variance of data. It is important to stress that these variance sources are man made and are consequently called “artefacts”.

Attempts to correct for some of the artefacts can be performed using additional information or assumptions like sample sizes, study means, standard deviations, estimates of reliability or other data. The foundation for this is the division of variance in one part influenced by the actual relations in nature and the part influenced by artefacts. By discovering the size of certain artefacts and knowing the ESs containing both quantities, the true, natural relation can be calculated. Since some artefacts can not be measured directly, attempts have been made to estimate them. However, some artefacts are hardly to be corrected at all. An additional problem is introduced as the amount of influence an artefact might have on a meta analysis might vary across studies. A way to compensate for this is to weight each study by the amount of artefacts it contains and only then to add it to the aggregation process (Hunter & Schmidt, 2004). For a conversion of these ideas for this work, see section 4.7.2.

3 Study Questions

After describing basic methods of meta analyses in chapter 2 and giving an overview of possible problems and their solutions, in this chapter the study questions of the thesis at hand are described. The general aim is to replicate the results Stayner et al. (2007) and to apply some advanced analyses to the given data. Then, after adding up the data and results, a range of possible overall results will be analysed through the CRM to show the influence of subjective decisions on the relation between ETS and the risk to come down with lung cancer. Hence the greater question is:

How *robust* are the given results as discovered by Stayner et al. (2007), can they be replicated by others, and if not, how much do they vary?

3.1 Replication Phases

A first replication phase will be based solely on the data and results reported in the original article. No data is collected directly from primary studies for this step. The study question here is:

(I) Are the results given in Stayner et al. (2007) to be replicated?

In the next step primary studies are included into the analysis. The same list of studies is collected and newly coded on the variables used in the original publication. The guiding question here:

(II) Is the replication of the results of Stayner et al. (2007) possible if the data has to be newly coded?

The same methods have to be used for this part as they are reported in the original study, the same specifications analysed. However, no results reported in Stayner's article are to be used to help with the analysis. It was tried to replicate implicit rules as well.

3.2 Study Questions Related to the CRM

As described above the CRM was developed to discover possible differences in results caused by subjective decisions made implicitly or explicitly by the conductor

Study Questions

of the meta analysis. Hence the resulting study question is:

(IIIa) Do the proposed crossroads have a significant impact on the results of the present meta analysis?

To be able to answer this question the range of results is explored, as it is set by the available data. One point is to find the absolute minimum point which can be produced out of the underlying data:

(IIIb) What is the minimum mean RR contained in the data and which way leads to this result?

Last but not least four “theories” about a reasonable way through the CRM, set in advance, are introduced adding up to the question:

(IIIc) What is the span between the results of the chosen theories?

3.3 Study Questions Related to the Advanced Analysis

Not being considered in the original study the study quality is added as a moderator to this thesis for further gain of information:

(IV) Does the quality rating of the primary studies explain a significant part of the data variation?

The last study question concerns the broad field of study artefacts, as they are introduced by Hunter and Schmidt (2004):

(V) Does the introduction of study artefacts shed additional light on the research question of ETS exposure at the work place and lung cancer?

The methods used to implement these questions are described in chapter 4, the results can be found in chapter 5. In the discussion section (see chapter 6) some critical views on these study questions will be given.

4 Methods

This section describes methods used for the thesis at hand. The implementation of Cooper's five step plan is described. Then the methods used in Stayner et al. (2009) are portrayed as closely as possible as well as methods for the operationalization of the CRM and the advanced analyses.

4.1 Problem Formulation and Data Collection

In section 2.1.1 the problem formulation was described as the first step taken in a meta analysis. The same procedure was assumed in this meta analysis. The major theme, to test the robustness of Stayner et al. (2007) results and to analyse the results for a possible range due to different crossroad decisions, was defined and operationalized through six study questions. These are described in chapter 3.

As explained above, for replication purpose *data collection* for this meta analysis was reduced to the mere search process for the 22 Studies used in Stayner et al. (2007). In order to receive these, a literature search was conducted in the University of Mannheim and the University of Heidelberg literature network. All articles were found through these sources. In the literature section the primary studies are marked with a * and they can be found in the Appendix, section E-c.

4.2 Evaluation of Data Points

For the evaluation of data points for this meta analysis two topics had to be accounted for. Firstly, every aspect reported in Stayner et al. (2007) had to be coded. Secondly, data for the CRM and the advanced analyses was needed. For this purpose a coding sheet and an according coding manual was written before the coding process was started. During the coding process of the primary studies several variables were added, others had to be discarded for various reasons.

4.2.1 Basis for Data Point Evaluation – Coding Sheet and Manual

Building the manual for coding required several steps. The data basis for this meta analysis had to describe several general themes and the content of the coding sheet was sorted into four blocks to satisfy this need. The content of these blocks is:

Methods

1. variables describing the key study attributes,
2. variables derived from the STROBE article (Vandenbrouke et al., 2007) to code for quality issues,
3. variables needed for the ES calculations and
4. variables originally developed for quality criteria in psychological analyses and adapted to the present study for health issues.

Example items for each block can be found in the next chapter, the complete sheet and manual in the appendix.

The coding manual was designed to describe every variable thoroughly, to illustrate possible values a variable can adopt and to assign distinctive eight digit codes for each variable. The according coding sheet was build in Microsoft Excel and contains the codes for each variable, its complete name and again its possible values. An overview of the coding content can be found in Figure 6. Here the number of originally designed items is displayed, during the later coding process not all items could be accounted for.

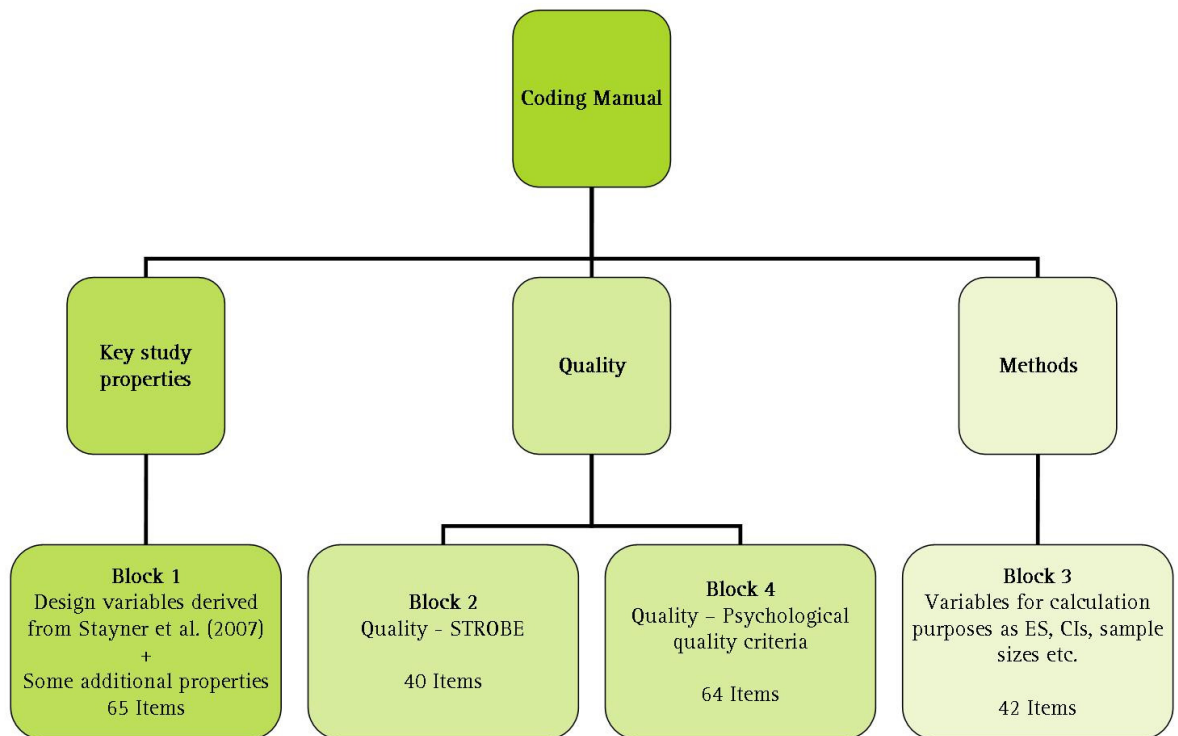


Figure 6 Overview of the coding basis for this thesis. For each block the number of items is reported.

4.2.2 Content of the Four Coding Blocks

In this section selected items are reported to give some insight into the data basis of the present meta analysis:

Block one describes key study features like the authors of the study, year of publication, overall number of cases and controls or forms of cancer included into the study. Accordingly, most variables in block 1 have a nominal scale.

Block two contains quality variables translated from the STROBE recommendations for high quality epidemiology publications (Vandenbrouke et al., 2007). These recommendations were translated to quality criteria by judging, whether each recommendation was met in the study or not. Converted into a binomial value of 1 for “recommendation sufficiently met” and 0 “recommendation not met” an overall quality rating can be determined by adding up all variables of block two. This way a maximum quality rating of 40 points could have been achieved in block two. Exemplary variables in this block are “The abstract explains the scientific

Methods

background and rationale for the investigation being reported?”, “Efforts to address potential sources of bias are described?” or “Reasons for non-participation for each stage are given?”. Of course here the coder had to rely on what was reported in the primary study, if there was a bias analysed but not reported a 0 was coded.

Block three consisted of closely coded information of data needed for various statistical calculations. This mainly concerns the ESs. Several new items for the CRM are to be found in this block as well. Example variables are the coding of the number of persons needed to calculate the RR or OR manually, just like the calculated CIs for each ES. For each value calculated the according value in the primary study was added when one was reported. Other variables were e.g. the ratio of cases and controls, significance calculations or the definition of intensity or duration for the ESs at hand. Variables in this block largely have a metric scale. Additionally this block varies most due to differences between single ESs (as long as a study reported more than one ES), whereas the other blocks mainly vary due to differences between studies.

Block four contains information about the quality of the primary studies, this time derived from quality criteria found in psychological studies. Examples are issues with external or construct validity or with reliability. As a rough basis information from the meta analysis published by Wittmann and Matt (1986) was used, in addition to artefact information as described in Hunter and Schmidt (2004). Most items in this block have an ordinal scale, however in order to be able to build a sum score as it has been done in block two many variables are dichotomised in a second step. This “second step” data was used for the later intercoder reliability calculations. After many items had to be removed, because the information to code them was not given, a maximum score of 10 could have been reached in this quality block. Example items are the reliability of the study at hand, whether there was a matching between cases and controls or whether there was a limitation of the external validity.

Especially in block four there were many variables which could not be coded because of varying reporting quality. These items are excluded from the coding

sheet but can be found in the coding manual marked as “raus”. Other blocks contained items not to be coded as well, only block two was used completely. After the reduction block one consisted of 60 items, block three of 30 and block four of 31 items.

4.2.3 Interrater Reliability

As described above, Krippendorff’s method to define the interrater reliability has established itself both in the social science and in other subjects (Bosnjak & Viechtbauer, in print). Values can vary between 0 and 1, the higher the results, the better the reliability. To calculate the Alpha in this thesis a SPSS macro program was used (Hayes & Krippendorff, 2007). Overall, three coders analyzed the primary studies, two are the authors of this thesis and Pahl (2009), another coder was trained through the manual. Several tests coding sessions were conducted until no further questions emerged. Only then the final coding phase took place. All coders were psychology students close to graduation (MSc).

Krippendorff’s Alpha was calculated through three randomly chosen primary studies which were coded by all three coders. Based on block 1, 2 and 4 a reliability of .84 was achieved. All three blocks were reduced to a nominal scale for the calculations. For block 3, containing the values for calculating the ESs, no interrater analysis was conducted. Instead all studies were coded by all coders and checked for mistakes when differences appeared.

After sufficient intercoder reliability was achieved the remaining 19 studies were split between the three coders. The dataset with the results for the intercoder reliability analysis can be found in the appendix G-c section.

4.3 Data Analysis and Interpretation; Presentation of Results

In the data analysis phase of the present meta analysis several steps were taken. At first all analysis methods used in Stayner et al. (2007) were tried to be identify. An analysis plan was developed based on this information. Then a first replication was started, executing each step of the analysis plan with only the data set derived from the original study of Stayner et al. (2007). All data from primary studies was

Methods

ignored. Methods of this part can be found in section 4.4, results in section 5.2.

For the next replication procedure all data from the original study was disregarded and the data obtained by the primary studies was used instead. Again it was tried to repeat each analysis step described in the original article. The procedures can be found in section 4.5 and the results in section 5.3.

The risk calculation of the first and the second replication varied. There is a difference between RRs as the ES value is called in Stayner et al. (2007) and ORs as used in the second replication. Both values depend on the same data, the content of a 2x2 contingency table (see Figure 7), however they use the content through two different equations.

	O1	O2	
G1	a	b	n_1
G2	c	d	n_2

Figure 7 Exemplary 2x2 contingency table as used for the calculation of relative risks or odd ratios. G possibly represents people having a disease vs. people not having the disease. O may stand for a kind of environment exposure present or not present. a to d stands for the (real) distribution of people belonging to one group or another. n are sums of the groups.

For the example of exposure to ETS and the risk of lung cancer four groups are possible. People who have cancer are called “cases” and people who serve as a control group (without lung cancer) are called “controls”: The first group contains the cases with exposure to ETS (a), the second controls with ETS exposure (c), the third cases without exposure to ETS (b), and finally there are the controls without exposure to ETS (d). In this example G1 would stand for “cases”, G2 for “controls”, O1 for “ETS at the workplace” and O2 for “no ETS at the workplace”. The allocation of G1, G2, O1, and O2 is somewhat arbitrary, which will become important when talking about the difference between relative risks and odd ratios.

The equation used for the calculation of RR is:

$$RR = \frac{a/n_1}{c/n_2} \quad (1)$$

Whereas the equation for OR is:

$$OR = \frac{ad}{bc} \quad (2)$$

In this thesis the OR is preferred, as it is independent from the order in the contingency table at hand, as opposed to the RR which changes dependent on the allocation of the labels in the table. If in one primary study the “G” is reported as the exposure and the O stands for the group of people, ORs still will lead to the same results, while RRs will not. The use of ORs is in agreement with Deeks (1998), who states that ORs are “the best estimates of relative risks that can be obtained” (S.1156).

Variance of ORs is calculated by the equation:

$$v = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d} \quad (3)$$

In general it is recommended to use logarithmized ORs as they achieve a closer approximation to a normal distributed variable and to get a symmetric variable. For this reason all ORs or RRs in this meta analysis are transformed into logarithmized values for calculations and back for the presentation of results.

After completion of the replication the CRM was tested, using data from both Stayner et al. (2007) and the primary studies. Due to time restrictions three crossroads were chosen and the resulting range identified. For a close description of the used methods see chapter 4.6, for the results see chapter 5.4.

Finally several advanced analyses were conducted, presented in chapter 4.7 and 5.5.

Concerning the presentation of results, the size of the meta analysis at hand made it necessary to split the content into two separate but related theses. Several different analyses and especially the analysis of the extreme maximum of the present analysis range can be found in Pahl (2009).

4.4 Replication Using Data from Stayner et al. (2007) – First Replication Phase

As described above, a replication using only the data reported in Stayner et al. (2007) was conducted. In this chapter the analysis steps and methods which were repeatable are described.

4.4.1 The Overall Effect Size

At first a mean overall ES has to be calculated using the primary study results as shown in Stayner et al. (2007) table 1. Both the fixed and the random ESs are reported. The value for ES calculations in the original article is the RR to come down with cancer when exposed to ETS at the workplace. These RRs were logarithmized for later calculations. To receive the residual variance logarithmized CIs were used:

$$v = \left(\frac{\ln(UC) - \ln(OR)}{1.96} \right)^2 \quad (4)$$

In this equation the UC stands for the upper CI, v stands for the variance. To calculate the weights for each ES the inverse variance was determined. As the CIs reported by Stayner et al. (2007) are partially not symmetric, not logarithmized and additionally rounded, it is important to describe other possibilities to determine the variance, e.g. through the upper and lower (LC) CI:

$$v = \frac{[\ln(UC) - \ln(LC)]^2}{(2 * 1.96)^2} \quad (5)$$

Additionally, equation (4) can be reformulated to use the lower CI to calculate the variance. Having the exact CIs all these methods would lead to the same results. However, through the problems named above (e.g. imprecise rounding), partially quite big differences result by using the different equations on the data set as reported by Stayner et al. (2007). To minimize this error only equation (4) was used in this thesis.

To calculate the mean ES the following equations were used. For fixed effects models the base assumption is that there is a true effect θ_i which is the underlying value for all primary studies (Mengersen, 1995; Schulze, 2004). By aggregating

Methods

primary studies this true value can be approximated. In addition an error term ε_i is introduced to the equation. Following classical test theory an observed value y_i consists of the true effect θ_i and the measurement error ε_i adding up to the equation:

$$y_i = \theta_i + \varepsilon_i \quad (6)$$

The i stands for each primary study. To achieve the estimated true value $\hat{\theta}$ the observed values y_i are weighted by w_i through the following equation:

$$\hat{\theta} = \frac{\sum w_i y_i}{\sum w_i} \quad (7)$$

The according variance is determined by equation (8):

$$\text{Var}[\hat{\theta}] = \frac{1}{(\sum w_i)} \quad (8)$$

As opposed to the assumption of homogeneity of variance for linear regressions or analyses of variance, here no such assumption can be made. Due to the often highly different sample sizes in meta analyses the sample variance will be different for each primary study (Viechtbauer, 2007b).

In order to be able to calculate a mean overall effect size, a weight for each primary study needs to be determined. Schulze (2004) recommends the use of the estimated inverse variances as these minimize the pooled variance of $\hat{\theta}$. w_i is calculated through the equation:

$$w_i = \frac{1}{v_i} \quad (9)$$

As Stayner et al. (2007) do not report their method of weighting we had to assume this method.

To estimate the mean ES with a random effects model an additional term τ_i is introduced to the equations named above (Lipsey & Wilson, 2001). As described in chapter 2.1.4 this term stands for a “true” random (residual) heterogeneity present in the study universe. That way no exact true effect can be approximated, but only a mean “true” ES μ_θ . The heterogeneity term for each primary study and the mean

true ES add up to the true population ES θ_i :

$$\theta_i = \mu_\theta + \tau_i \quad (10)$$

Based on equation (6) and (10) in a random effects model the observed value consist of:

$$y_i = \mu_\theta + \tau_i + \varepsilon_i \quad (11)$$

The true mean ES can be estimated through the equation:

$$\hat{\mu} = \frac{\sum w_i y_i}{\sum w_i} \quad (12)$$

The variance of this value can be calculated through:

$$Var(\hat{\mu}) = \frac{1}{\sum w_i} \quad (13)$$

And finally the weight is influenced by the random heterogeneity:

$$w_i = \frac{1}{v_i + \hat{\tau}^2} \quad (14)$$

It is assumed that the ESs of the primary studies vary normally distributed around the true mean ES (Viechtbauer, 2007a). For a discussion on fixed and random models, see chapter 2.1.4.

A plot of the primary study ESs and the according CIs is shown in the first figure of Stayner et al. (2007). This plot, commonly referred to as Forest Plot (Lewis & Clarke, 2001), can be sorted by the size of the ES to enhance the comparability of each study to another and the mean ES (see Figure 12). In this version e.g. possible differences of the CI size depending on the ES can be recognized more easily.

The mean ESs were calculated through a SPSS macro designed by David B. Wilson. To determine the CIs for the mean ESs SPSS uses the normal distribution. Some difficulties with this approach might be found as not all statistic analysis software uses the same base (see 6.2). Homogeneity tests were applied on the calculated mean ESs. Stayner et al. (2007) used to different tests, the DerSimonian and Laird

test (1986) and the Likelihood Ratio Test. As no other test was available for this thesis the Q-test provided by SPSS was used:

$$Q = \sum_{i=1}^k w_i (y_i - \hat{\theta})^2 \quad (15)$$

This homogeneity test is based on a χ^2 distribution with the degrees of freedom of $df = k - 1$, k representing the number of primary studies. Q is the sum of the squared difference of the observed ES y_i and the estimated true mean ES $\hat{\theta}$. Each squared difference gets weighted by the inverse study variance w_i (Hedges & Vevea, 1998). A significant result can either be an indicator for a moderator variable or a random unique study variation. However, the counter-argument that no moderator analysis or no random model is needed when the test is not significant must be handled with care, as the test lacks statistical power (Viechtbauer, 2007b).

Furthermore Stayner et al. (2007) conducted a moderator analysis, testing chosen key study designs features on their effects on the overall result. As listed in the first table of the original article these features are: The reference, the location, the time period, the gender, several covariate adjustments, the exposure period and histological confirmation. How these features were tested and which of them exactly were included is reported somewhat unclear, we decided to test the features as listed above. How the influence of the author was tested is not reported, in this thesis an approach where all author related to one another are grouped and tested against each other was used. For the first replication phase this results in a dummy coding of studies related to either Boffetta or Kabat, with studies related to non of them serving as a comparison group. The other key study feature moderators were either dummy coded (e.g. gender) or added individually into a moderator analysis. Both fixed and random effects models were tested.

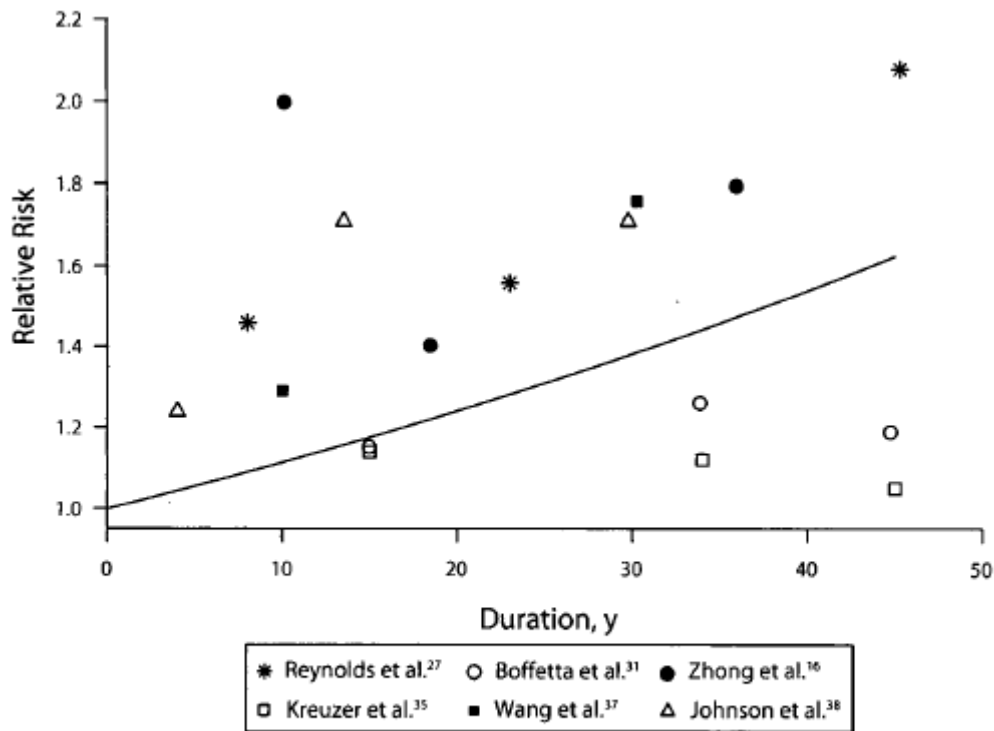
4.4.2 Exposure Response Analyses

In Stayner et al. (2007) a special form of moderator analysis is called *exposure response analysis*. Under this caption both the intensity and the duration of ETS at the workplace were tested separately. The number of primary studies was reduced

Methods

by this process, as only few studies reported details needed for these analyses. In the case of the duration the ES of each subgroup reported was extracted and included into a meta regression, in the case of intensity simply the highest intensity ESs were aggregated to a new mean ES like described in section 4.4.1; both fixed and random effects model were calculated. For the first analysis six studies were available containing the necessary data (Boffetta et al., 1998; Johnson et al., 2001; Kreuzer et al., 2000; Reynolds et al., 1996; Wang et al., 2000 and Zhong et al., 1999) whereas seven studies were included into the highest intensity aggregate (Boffetta et al., 1998; Johnson et al., 2001; Kabat et al., 1984; Kalandidi et al., 1990; Kreuzer et al., 2000; Lee et al., 2000 and Zhong et al., 1999).

The advantage of the first analysis is the prediction equation. For this purpose studies which reported two or more additional ESs at different duration times were included into a linear meta regression. In the end the duration of exposure can predict the risk of coming down with lung cancer with a certain amount of accuracy. As the values of the data points are not reported in Stayner et al. (2007) some compromises had to be made to achieve at least an approximate replication. In order to estimate the data points used for the meta regression the coordinates of the points shown in the second figure of Stayner et al. (2007) were read out (see Figure 8). To add accuracy the program "g3data" (Frantz, 2000) was used. The figure is moved into the program, the points to be measured are selected manually and the program translates these points into coordinates (see Appendix F-b.). No replication of the weighting was possible, as literally no information was given on this topic.



Note. The diagonal line is the fitted fixed model effect of duration.

Figure 8 Duration exposure response analysis as displayed in Stayner et al. (2007), second figure. Points displayed in this graph are the base for the direct replication in this thesis.

To perform the meta regression SPSS was used. In a first step based on the coordinated obtained by g3data and Stayner et al. (2007) second figure were transferred into a b_k value with an according standard deviation v_k . Stayner and colleagues report to have used an article by Greenland and Longnecker (1992) as a base for the methods, however this article is only partially applicable for the first replication phase, the database was not sufficient. No correction for dependent data points could be accomplished; however the equations for the estimated pooled b_p and the applicable standard error s_p were used for the further replication:

$$b_p = \frac{(\sum_k b_k / v_k)}{(\sum_k 1/v_k)} \quad (16)$$

and

$$s_p = \sqrt{\sum_k \frac{1}{v_k}} \quad (17)$$

Methods

Variation could not be calculated for all studies (mainly Kreuzer, 2000). In consequence v_k was fixed to 0.0001 for this thesis, an approximation to the mean v_k , which only affects the pooled standard error, not the pooled b-value. After determining the needed values and achieving the pooled b-value using equation 16 a predictor equation can be build as follows:

$$\hat{OR} = 45 * b_p \quad (18)$$

A CI for this predicted value is calculated through the following equation:

$$CI_b = \exp[45 * (b_p \pm t_{krit,df} * s_p)] \quad (19)$$

Homogeneity can be tested using the Q-test as described above (see equation 15), by pooling the b_p values with the fixed effects.

4.4.3 Sensitivity Analyses

The *leave-one-out* approach was mainly used as sensitivity analysis in Stayner et al. (2007), checking the influence of each primary study on the overall ES by removing it from the calculation. It was not reported whether each ES (25) or each study (22) was removed. Correspondingly both possibilities were tested during the replication. This approach was used for the exposure response analyses as well, resulting in 7 (6) overall effects with each one omitting a primary ES (study) for the intensity moderator and 6 for the duration moderator. In addition to the leave-one-out analyses the effect of those studies was tested, which had any kind of adjustment done as compared to the studies without adjustments.

4.4.4 Publication Bias

The effect of a bias caused by missing or not published journals, was tested using a funnel plot. Stayner et al. (2007) used this graphical approach by plotting the RR of each study against the inverse variance of the logarithmized RRs. The original plot can be found in Figure 14, here compared to the replication of the first phase.

4.5 Replication Using Data from the Primary Studies – Second Replication Phase

The second replication stage is aiming at the same analysis steps like the first stage. However, this time only information contained in the primary studies is used. Data derived from Stayner et al. (2007) was completely ignored. Only analysis steps different to chapter 4.4 are described in this chapter. The major aim of the replication based on the primary studies is to achieve the same data and to conduct the same analyses as described in Stayner et al. (2007) in order to compare the results of the original and the replication. The second replication phase gets somewhat closer to the original idea of replications.

4.5.1 The Overall Effect Size

Whenever the raw data was reported in the primary study, in addition to the adjusted ES the OR was calculated using equation 2. No data was available to reconstruct the adjustments many primary studies did (for an overview see Table 2). Differences between reported primary data and data reported in this thesis will partially be due to this fact. In many cases separate analyses of male and female cases were possible. Several times only a female number and an “all cases” number were reported, the necessary raw data for men was achieved through simple subtractions. Additionally a problem came up concerning the definition of ETS at the workplace (see section 4.6.1). This problem and the description of a high number of subgroups e.g. describing different kinds of intensity or lengths of duration, lead to a high number of available ESs for several primary studies. A single adjusted ES with its CI was hardly coded for any primary study, complicating the choice of “the correct” ES for the replication.

In order to find this “correct” ES the following rules were highlighted based on Stayner et al. (2007). Some rules are explicitly reported in the original study, some are reported but not followed and some are of a more implicit kind:

1. If there is a reported adjusted ES it is to be used.
2. If a separation of male and female cases is possible, this was done.
3. Other subgroups besides gender have to be ignored.

4. The second definition of ETS exposure group was used when available (see section 4.6.1).

Methods to calculate the mean ES and the test for homogeneity have been the same like in the first replication phase. Key study design features were tested as well. The list of features was slightly different, compared to the original study. The influence of authors was tested by separating the authors in four groups, related either to Boffetta, Wu, Kabat or non of them. Several adjustments reported in the primary studies and not dealt with in the original study were added (e.g. education or consumption of vegetables), the variable “exposure period” on the other hand was dropped, as no reliable information about this topic was given in the primary studies. Beside these changes the same procedure was used to test the key study design features. All methods of this part are described in section 4.4.1.

4.5.2 Exposure Response Analyses

In order to replicate the intensity analysis reported in Stayner et al. (2007) the ESs related to the highest intensity were collected. Since some primary studies report several subgroup data which could be classified as an intensity measure, the rule “cumulative exposure rather than intensity of exposure” (Stayner et al., 2007, p. 546), was introduced in the original article as well as in this thesis. This is an additional rule to the ones described in section 4.5.1. Homogeneity was tested through the Q test.

To conduct the meta regression all data concerning lung cancer RRs and the duration of ETS exposure were accumulated. Most of the time different time intervals are reported in the primary studies. Stayner et al. (2007) claim they have used the middle of these intervals to conduct the regression. An exception was made when the last interval was open ended (e.g. exposure more than 20 years), here this last limit was multiplied with 1.5. When the value would exceed 45 years of exposure, a cut was made, no exposure time endpoint bigger than 45 entered the calculations. In the second replication phase a correction of correlating (dependent) regression points as described in Greenland and Longnecker (1992) was possible,

since the primary studies reported most of the necessary data. The set of rules which ESs to take when multiple ESs were available changed only for the gender: This time ESs for men and women combined had to be chosen above the other possibilities (see . 4.5.1 for the other rules). Concerning the time intervals the primary study of Kreuzer et al. (2000) is somewhat of an exception to the other studies. Here the interval is reported in hours, not in years. Regarding that the study was conducted in Germany (and its local employment law) these hours had to be recalculated into years in order to add the study to the regression. Based on the assumptions that a workday will have been 8 h long, a week will have contained 5 work days and the year will have contained about 46 weeks of work (30 days are off because of holyday time) a value of $46 \cdot 5 \cdot 8 = 1840$ hours of work a year was calculated. Now the hours reported in Kreuzer et al. (2000) were divided by 1840 to receive an approximation of the years of ETS exposure at work. Even though this procedure was not reported in Stayner et al. (2007) this seemed to be the only possibility to add the study to the regression.

Kreuzer et al. (2000) stands out for another feature. Normally comparison groups are not exposed to ETS at all. However, in the subgroup calculation reported in Kreuzer et al. (2000) a group possibly exposed up to 29000 h of ETS is chosen as a reference group. As the size of the group not exposed at all is reported in this study, that value was chosen above the value reported in the according subgroup analysis (see table 4, p. 245, Kreuzer et al., 2000). The only drawback of this method is that adjusted values had to be rejected that way, as the ORs included into the meta regression now had to be calculated from the raw data.

In the next step a SAS macro (Zack 1996) was used to accomplish the correction for correlated data points in a regression (see appendix F-b for information on the macro). Besides this correction the further analyses was conducted the same way as in the first replication phase (see section 4.4.2.).

4.5.3 Sensitivity Analyses

Here no differences appeared as compared to the first replication phase, the leave-

one-out analyses and the control for adjusted vs. unadjusted ESs were conducted in the same way like described in section 4.4.3.

4.5.4 Publication Bias

The same funnel plot was build with the newly coded data as compared to the data derived from Stayner et al. (2007). However, a third version was made as a recommendation how to enhance the probabilities of a funnel plot (Sterne & Egger, 2001): Instead of using the inverse variance of the RR like in the original study, the standard error was plotted. To provide a comparison both plots will be displayed in the method section (see Figure 14 and Figure 18). Additionally, here both plots use logarithmized ESs to make them comparable.

4.6 A Crossroad Model

As described above, in the CRM subjective decision options are to be tested on their effect on the results of this meta analysis. This section describes, which crossroads were chosen to be analyzed in this thesis; these will be described in more detail.

4.6.1 Description of Crossroads

The first crossroad describes the decision whether *only adjusted values* or *only unadjusted values* should be entered into the meta analysis (for a discussion on why these should not be mixed see section 2.1.7). Stayner et al. (2007) did report a mix of both adjusted and unadjusted values and an analysis which accounted only for the adjusted ESs (resulting in an aggregation of only 16 ESs). The analysis of only unadjusted values was missing.

The second crossroad analyzed in this thesis concerns the *gender of the cases and controls* presented in the primary studies. Some studies have a strictly female sample, some have a mixed sample. Depending on the generalization to men or women or cases in general a highly different number of ESs are available, producing possibly different results.

The third crossroad concerns the definition of ETS exposure at work and thereby the very heart of the question whether ETS exposure is harmful or not. The study

sample used by Stayner et al. (2007) contains several different definitions of what exactly “exposure at work” really means. Since most studies lack an exact definition, the questions cases and controls were asked provide some evidence towards which definition was used. Only one study explicitly dealt with this problem (Koo, 1984), Figure 9 is based on the idea displayed in the first figure of this study.

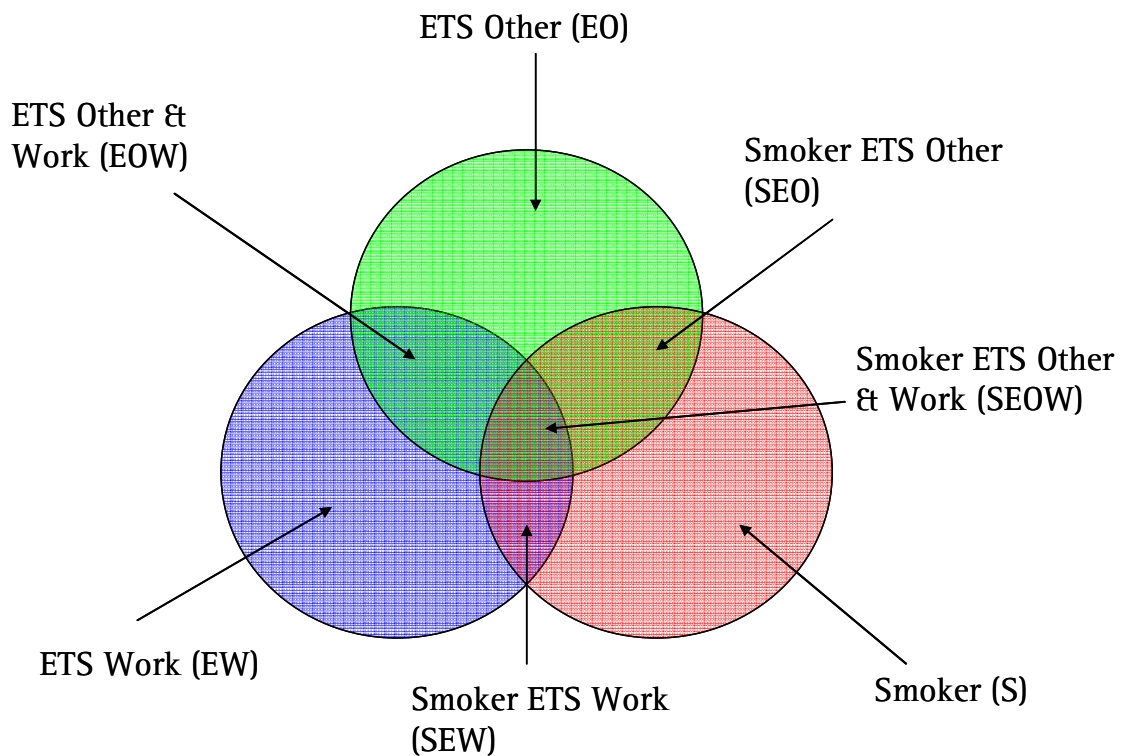


Figure 9 Different possible groups of ETS exposure.

In Figure 9 three different potential kinds of smoke exposure are drawn. Of course smokers are exposed to cigarette smoke but are to be excluded from this study (marked with S). Then there can be cases exposed to ETS at a different place than at work. This is called ETS other (EO) in the figure above. Although these people are exposed to smoke without being smokers themselves, only to be exposed at work is of interest for this thesis as it was for the original study Stayner et al. (2007). Now for the issue that is of interest: The circles do have intersections. Active smokers can be exposed to ETS at other places (SEO) and employees exposed at work can be additionally exposed at home (EOW). It goes without saying that all intersections

Methods

related to smokers are to be excluded. However, in the provided definition “We sought to quantitatively evaluate the association between workplace environmental tobacco smoke exposure and lung cancer” (Stayner, 2007, p. 545), all groups exposed at work *and* at other places can still be part of the sample.

Three differently rigorous definitions can be constructed out of the areas displayed in Figure 9: the first and most rigorous will accept only and exclusively people who were exposed at work. The cases and controls exposed to ETS at home or other places would have to be excluded. This would be the EW field only. The second, somewhat more tolerant definition would accept cases exposed at work who might be exposed at home as well. This is represented by the fields EW + EOW. Finally, the third and most lax definition would allow people exposed only at home, as long as they are no smokers themselves. This results in the area EW + EOW + EO. Some primary studies provided small hints, which definition they used. If cases and controls e.g. were asked if they were “exposed at the workplace and only at the workplace?” this would be an indication for the rigorous definition, whereas “were you exposed at the workplace?” can only match the second definition. Here, cases and controls exposed at other places as well might still answer “yes”. A question related to the lax definition would be “were you exposed to ETS?”. As in the group positive to this question still cases exposed at work are present it does fit to the study question (and was included in the original study, see discussion). However, persons e.g. exposed only at home might be included as well.

Naturally, these three crossroads chosen for analysis are only a small assortment of the ones which are imaginable, derived only from the evaluation of data points phase of a meta analysis. For a number of further suggestions see Figure 10.

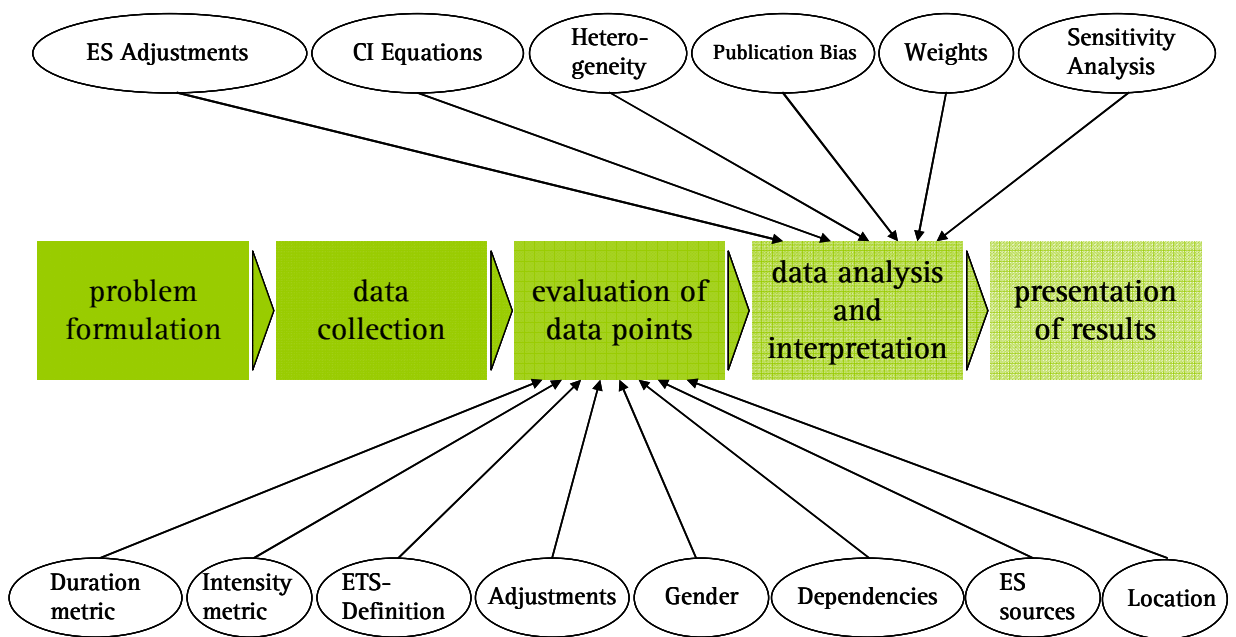


Figure 10 Display of a number of possible crossroads at evaluation of data stage and data analysis stage.

Promising crossroads for the evaluation of data points stage are the choice of a duration metric, the handling of dependent values, or the detail computing level like in “location”. Some examples of the analysis stage are the choice of the statistical computing package, different equation use for CI or ES calculation or a different estimator base for the homogeneity analysis. Here only several possibilities of the third and fourth step of a meta analysis are shown. However crossroads are imaginable at each stage of a meta analysis and even for the two displayed stages many more crossroads are possible. E.g. Abrami et al. (1988, p. 160) state that:

“Effect magnitudes (sizes) are computed according to procedures described by Glass et al. (1981) and others (e.g., Rosenthal, 1984) to transform the findings from different studies to a common metric. [...] But the calculation of individual study outcomes is often not straightforward and may involve a series of decisions that can affect the value of each effect magnitude[...].”

Admittedly the number of ESs usable for analysis varies stronger the more crossroads are added to the model. An extreme example would be a chosen path of “only unadjusted ESs” + “women only” + “rigorous definition”, here only three of 22 studies report sufficient data to perform the analysis, though none of the paths

chosen are wrong. In order to provide a calculable crossroad model, the three crossroads described above were chosen. They were reported clearly in most primary studies and all studies could be included into the analysis at some point. Last but not least the three crossroads were chosen for their effect on the study question, as the inclusion of different exposure groups should have an impact on results, men and women have highly different prevalence in the sample in addition to the probability of coming down with lung cancer and the adjustment is focus of an ongoing discussion in meta analytical controversies (see section 2.1.7).

After testing for the influence of the crossroads in a CRM it would be interesting to expose the extreme values possible due to a chosen path, and which decisions lead to them. This topic will be dealt with in the next section.

4.6.2 Finding the Upper and Lower Limits

Two ways are chosen to show the limits set by the CRM. The first follows rather blindly one or another direction, one could call it the “bad guy” or the “good guy” analysis (representative for a minimum mean OR result and a maximum mean OR result, respectively). This approach is blind to any theoretical framework developed in advance, as it just chooses the smallest or biggest ESs available, independently from the subgroup (e.g. different intensity exposure groups) or any crossroads. That way the minimum and the maximum mean OR possible will be found while using all primary studies. This thesis mainly deals with the calculation of a bad guy analysis. For comparison reasons the good guy analysis will be printed as well. The major approach concerning the good guy analysis can be found in Pahl (2009).

The second way includes theories specified in advance, as they might appear in a study field as lung cancer and the exposure to ETS at work. These theories set restrictions for the choice of ESs, not all primary studies might be included. Nevertheless one restriction might be to use as many ESs as possible. These theories will be described in more detail below.

The first theory is:” Only conclusions about women will be drawn”. Hence here only ESs concerning women are considered. Concerning the other crossroads, only the

second definition of ETS exposure groups will be used and the adjustment crossroad is regarded by calculating a value for adjusted and another for unadjusted ESs.

The second theory is related to the first one, only that it is rewritten for male cases and controls. Besides the gender differences the procedure was identical, the same rules applied and the results were separated according to the presence of adjustments.

The third theory takes a rigid perspective; it follows the study question stated in Stayner et al. (2007) exceedingly close: "Does ETS exposure at the workplace and *only* at the workplace have an effect on RR of lung cancer?". Here only persons fitting into the rigorous definition of ETS exposure will be allowed into the analysis. Besides that the other crossroads are flexible. Men and women can enter the analysis, though separated ESs are chosen above combined ESs, and adjusted and unadjusted values can enter the analysis, though adjusted values were preferred.

The fourth theory eventually concerns a method opposed to the third theory, as here the lax ETS group definition stood in the centre of interest. Besides this difference the process is repeated like in theory three, the same rules apply, the same position of points is used for the other crossroads.

4.7 Advanced Analyses

Without claiming to address all possible advanced methods available, some methods not used in Stayner et al. (2007) shall be introduced in this chapter, which might yield more information about the relationship of ETS exposure at work and lung cancer.

4.7.1 Quality of Primary Studies

One important, supposedly disregarded, point was the quality of the studies included into the meta analysis. To address this issue block two and four were introduced into the coding process, containing quality information based on STROBE criteria (Vandenbroucke et al., 2007)) and on validity information (Wittmann & Matt, 1986, Hunter & Schmidt 2004). This thesis will mainly

concentrate on the second quality aggregate, for a close analysis of the STROBE items see Pahl (2009).

Though a high number of items were included into block four, only a small number of these were eligible for analysis due to lacking or differentiating reporting practices in the primary studies. These eligible items had to be dichotomized to make them comparable to the block two quality items. Out of the 64 original items designated for block 4 only 10 were included into the final quality rating, which were the following items (the original German expressions are translated into English by the author of the present thesis):

“Equivalence” (04aequB4), “Presence of matching of cases and controls” (05mageB4), “Presence of raw data for ES calculations” (07esanB4), “Confirmation methods for cancer type described” (12metkB4), “Size of the catchment area of the present study” (24regcB4), “Interpretation accuracy in the control group” (29isgeB4), “Direct questioning of cases only” (30krkbB4), “Definition of Never-smokers given” (31deniB4), “Variability constraints in the UV” (36vushB4) and “Variability constraints of the sample” (38vastB4). For a closer description of the items see the coding manual in appendix G.

These items were introduced as a continuous moderator into the analysis after a sum was calculated. A high value would indicate a high quality whereas a small value stands for a poor quality. A maximum of 10 points could be reached as opposed to a minimum of 0 points for a very poor quality.

4.7.2 Correction of Study Artefacts: An Approach by Hunter and Schmidt

In this chapter the adaptation of the approach by Hunter and Schmidt (2004) to correct for study artefacts is described. In their book they discussed several possible artefacts and their influence on a meta analysis. So far this method was mainly used in study areas related to the field of industrial and organizational psychology, where it dealt with correlations. Now it must be reinterpreted for a health matter and ORs serving as ES. The so called “attenuation model” describes in Hunter and Schmidt the parts of an observed ES. This observed ES can be

influenced by the true effect *and* systematic or unsystematic artefacts, which decrease the observed relationships. The true effect and its CIs will be underestimated, if the artefacts are not accounted for. To solve this problem Hunter and Schmidt propose the introduction of the disattenuation model, through which the bias introduced by artefacts is corrected. This correction affects both the ESs of the study and the weight it will be given in the aggregation process.

Which of the proposed artefacts can be adopted to the meta analysis at hand is discussed below:

1. The sampling error describes a randomly varying difference of the population value and the study validity. This artefact can not be corrected in a single ES. However, as its size is mainly influenced by the sample size of the according study, it is accounted for to some extent by the weights, since these depend on the sample size.
2. Error of measurement in the dependent and in the independent variable describes an unsystematic reduction of the effect. Hunter and Schmidt propose a correction of each ES entering the study sample. The reliability for each instrument used for the concerning ESs is needed for this process, however neither the primary studies nor the original study Stayner et al. (2007) reported the reliability of the measurements they used. An attempt to judge the reliability of each instrument used in the primary studies ourselves failed as not enough information was published. Instead, a constant value was used, one for the dependent, one for the independent variable. An overall reliability estimation of a medical scientist was adopted from Ashton (2000), reliability of patient statements to their passive smoking status was assessed in Pron et al. (1988). Other studies came to different results than these two examples (e.g. Huang, 2008), however they still offer a sufficient base for this correction.
3. The dichotomization of continuous variables results in an artificial confinement of range, which in turn leads to a smaller ES. This can happen both on the dependent and the independent variable side. The adaption of

this artefact to the study sample at hand is not possible, as no artificial dichotomizing appeared. Besides, homogenous measurement approaches are needed to implement the correction for this artefact, which is not given in this study sample.

4. Artificial range variation can lead to an over- and underestimation of the effect at hand. The available data was checked for artificial restrictions. On the patients side the inclusion criteria varied depending on the definition of a never smoker, being sometimes more, sometimes less restrictive. However this information was not given often enough to correct for this bias, only 10 out of 22 studies reported their never smoker definition. A similar effect on the dependent variable side is called "attrition artefact". Here processes during the data collection of the primary study lead to a decrease of persons in the study sample. Information about in-patient deaths occurring during the assessment or patients who had checked out for some reason might have led to an successful correction of this bias, however this information was not reported once in the given study sample.
5. Deviation from perfect construct validity can appear in the dependent and independent variable. Opposed to the reliability artefact this is a systematic effect, based on the operationalization of the study question. The correction procedure is similar to the reliability correction, however exact knowledge of the study question at hand is needed to be able to asses the amount of deviation. In our research team no medical specialist was present, thereby this artefact was dropped. In future analyses we hope that collaboration with a medical scientist can be established to enable this correction. Partially this artefact is analysed by the introduction of quality as a moderator (see 2.5.1), as lacking construct validity can be interpreted as a quality problem.
6. Reporting and transcriptional error will appear probably in every published article. In section 2.2.4 some of these errors appeared in Stayner et al. (2007) are described, however it is not possible to estimate an overall error rate based on this information. As Hunter and Schmidt (2004) describe in their

book a correction of this artefact is near to impossible. A literature research for general error rates in scientific publications was not successful. Therefore this artefact was not accounted for in this thesis.

7. The last artefact describes extraneous variables which influence the results of a meta analyses. Effort was deployed to discover all possible moderator variables for the study question at had (see coding sheet and manual). Most of them could not be accounted for as the primary studies did not report information about them. Implementation of this artefact could be greatly enhanced when documentation practices of primary studies would improve.

For the reasons stated above so far only the correction for reliability is practical. Corrections have been applied to the independent and the dependent variable. As the work of Hunter and Schmidt draws mainly upon correlations the methods had to be adapted to suit a RR and OR related ES group. For this reason ORs used for this analysis were computed into correlations, then the artefact corrections were applied and in a last step the corrected correlations had to be recalculated to ORs.

The rather complex correction procedure was carried out through the following steps:

- 1) Conversion of OR values to uncorrected correlations:

$$r_{0i} = \cos \left(\frac{180^\circ}{1 + \sqrt{\frac{ad}{bc}}} \right) \quad (20)$$

- 2) Weighting with N was accomplished through:

$$\bar{r}_0 = \frac{\sum N_i * r_{0i}}{\sum N_i} \quad (21)$$

- 3) The uncorrected sampling error can be estimated through:

$$Var(e_{0i}) = \frac{(1 - \bar{r}_0^2)^2}{N_i - 1} \quad (22)$$

Methods

4) The identification of the disattenuation factor a through the dependent and independent variable reliabilities:

$$a = \sqrt{r_{xx} * r_{yy}} \quad (23)$$

5) The identification of the attenuation factor A :

$$A = \frac{r_o}{r_c} \quad (24)$$

6) The calculation of the corrected correlations for each primary study:

$$r_{Ci} = \frac{r_{0i}}{a} \quad (25)$$

7) The calculation of the corrected sampling error for each primary study:

$$Var(e_{Ci}) = \frac{Var(e_0)}{A^2} \quad (26)$$

8) In the weighting process through w_i not only the study size is accentuated for, but the size of the attenuation factor, as well:

$$w_i = N_i A_i^2 \quad (27)$$

On this database the weighted and corrected meta analysis parameter can be estimated:

9) The corrected overall effect (random effects model):

$$\bar{r}_c = \frac{\sum w_i * r_{Ci}}{\sum w_i} \quad (28)$$

10) The according corrected CI. k stands for the number of used studies:

$$CI = \bar{r}_c \pm 1.96 \sqrt{\frac{Var(r_c)}{k}} \quad (29)$$

11) The weighted and corrected variance of the correlations:

$$Var(r_C) = \frac{\sum w_i [r_{C_i} - \bar{r}_C]^2}{\sum w_i} \quad (30)$$

12) And finally the homogeneity estimation through the mean corrected sampling error, which in turn is divided by $Var(r_C)$ for the final estimation:

$$Ave(Var(e_{C_i})) = \frac{\sum w_i * Var(e_{C_i})}{\sum w_i} \quad (31)$$

13) The resulting meta correlation parameters can be reformulated as ORs through:

$$OR_C = \left(\frac{180^\circ}{\cos^{-1}(\bar{r}_C)} - 1 \right)^2 \quad (32)$$

A correction of the results Stayner et al. (2007) reported can not be accomplished (replication phase one) as no overall sample size is given. In the original study only case numbers are reported. The ratio of cases and controls differs significantly between primary studies. Some studies have a 1:1 ratio, some others ratios of 2:1 or 3:1. Lee (1986) even has a ratio of 1:17. Thereby an analysis based only on case numbers can not be recommended. Here primary studies would be given the wrong weight in the analysis.

The study sample of Stayner et al. (2007) meta analysis holds another problem. Five primary studies do not report their own sample size (Rapiti et al., (1999); Shimizu et al., (1988); Sun et al., (1996); Wang et al., (2000) and Wu et al., (1985)). Stayner et al. solved this problem by performing the weighting directly, using the reported variances of the concerning studies. However, this approach is not applicable for the Hunter and Schmidt approach, as the weights consist of the sample size *and* the attenuations factor (see equations 27 and 28). In the following these five studies had to be removed for this reason. Instead, the resulting artefact corrected meta analysis was based on 17 primary studies and was compared to an accordingly adapted meta analysis with 17 studies which were not corrected. The results of this analysis can be found in section 5.5.2.

Methods

Finally a rule of thumb needs to be brought up. Hunter and Schmidt introduced a 75% rule for homogeneity calculations. If at least 75% of variance in the corrected observed ESs is explained by the sampling error variance, homogeneity of the study sample can be assumed. The last 25% can be attributed to artefacts not yet discovered and are probably unsystematic. Hunter and Schmidt assume the use of a significance test for homogeneity as well, however they did not recommend it, as it lacks power (Hunter & Schmidt 2004).

5 Results

Through the following sections the results of this thesis are described. First the study sample and the resulting data set is described. Next the results of the meta analysis are presented, the replications, crossroad model and advanced analyses will be covered.

5.1 Data Basis

All in all more than 200 items were coded. To give an overview only a choice of items were described in chapter 4.2.2. For a complete description see the coding manual provided on the appendix CD.

22 primary studies are used for the study at hand, the same which were used in Stayner et al. (2007). These studies were published between 1984 and 2001 in several different journals. All studies are written in English. Concerning the gender ratio an apparent overbalance is present. 13 of the 22 studies have only female cases and controls. Of the remaining nine studies which do report results for women and men the number of women is much higher. E.g. Boffetta et al. (1999) included 4 male cases, compared to 66 female cases, Kabat et al. (1984) report a ratio of 1:2.6 men to women. No study of the present sample includes only men.

A special issue arises concerning the sample of Kreuzer et al. (2000). Some of their cases and controls are supposed to be part of the study of Boffetta et al. (1998), this is reported in Stayner et al. (2007). As both studies are part of the meta analysis these persons were excluded from the sample Kreuzer et al. provides. However, Stayner et al. do not report the resulting new case and control numbers, and neither Boffetta et al. nor Kreuzer et al. report to have had partially the same sample. Kreuzer et al. had to be included into this thesis without any adaptations.

5.2 Replication Using Data from Stayner et al. (2007) – First Replication Phase

The results are reported in the same order as they were in the original study in order to provide a maximum of comparability. As done in Stayner et al. (2007) first the overall results are reported, then the key feature analyses. Next the exposure-

response analyses are dealt with in addition to the sensitivity analyses. Finally the publication bias replication is displayed. All data used can be found in appendix A.

5.2.1 The Overall Effect Size

The mean ES is based on 22 primary studies and the resulting 25 ESs reported in the first table of Stayner et al. (2007). Replicating the fixed effects model an overall mean of 1.28 (CI: 1.17; 1.40) is found. The random effects model results in an overall mean ES of 1.27 (CI: 1.16; 1.40). The minor differences between the original values and our replication are probably due to rounding differences, as the table of Stayner and colleagues only presents values rounded to the second decimal digit. Differences on CIs possibly arise through the use of different statistical programs, as described above. The Q-test shows no evidence for heterogeneity ($Q = 26.23$, $df = 24$, $p = .34$), the random variance is $\tau^2 = .0050$.

All in all 6 key study design features are found to be significant, both using a fixed effects and a random effects model with an α of 5%. These moderators are: The time period ($Q = 7.11$, $p = .01$; 09bestB1), the number of cases ($Q = 4.44$, $p = .04$; 15nostB1), adjustment for diet ($Q = 4.59$, $p = .03$; dummy coded), adjustment for ETS exposure of smoking spouse ($Q = 4.59$, $p = .03$; dummy coded), adjustment for occupational exposure to other carcinogens ($Q = 5.27$, $p = .02$; dummy coded) and the exposure period (fixed: 4.07, $p = .0436$; mixed 3.88, $p = .0488$; dummy coded). Besides the exposure period the other key study design features analysed have literally no difference between fixed and mixed effects model. The results are transferred to a forest plot (see Figure 11).

Discussion

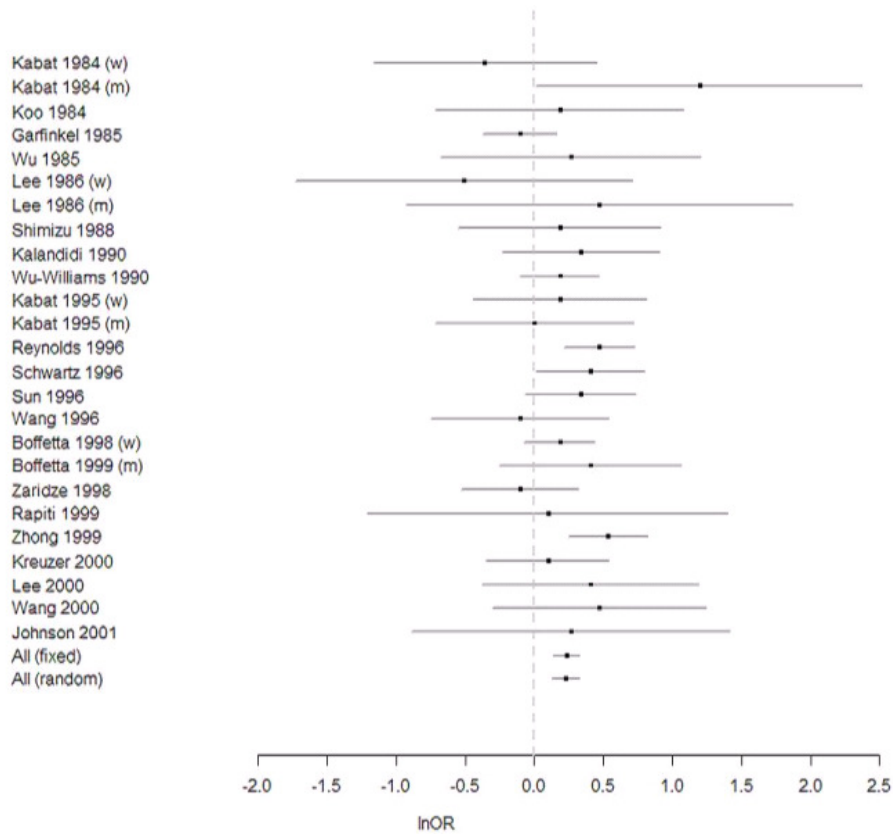


Figure 11 Forest plot first replication phase.

For comparison reasons the original plot is displayed in Figure 12, in addition to a plot which is sorted by ES magnitude.

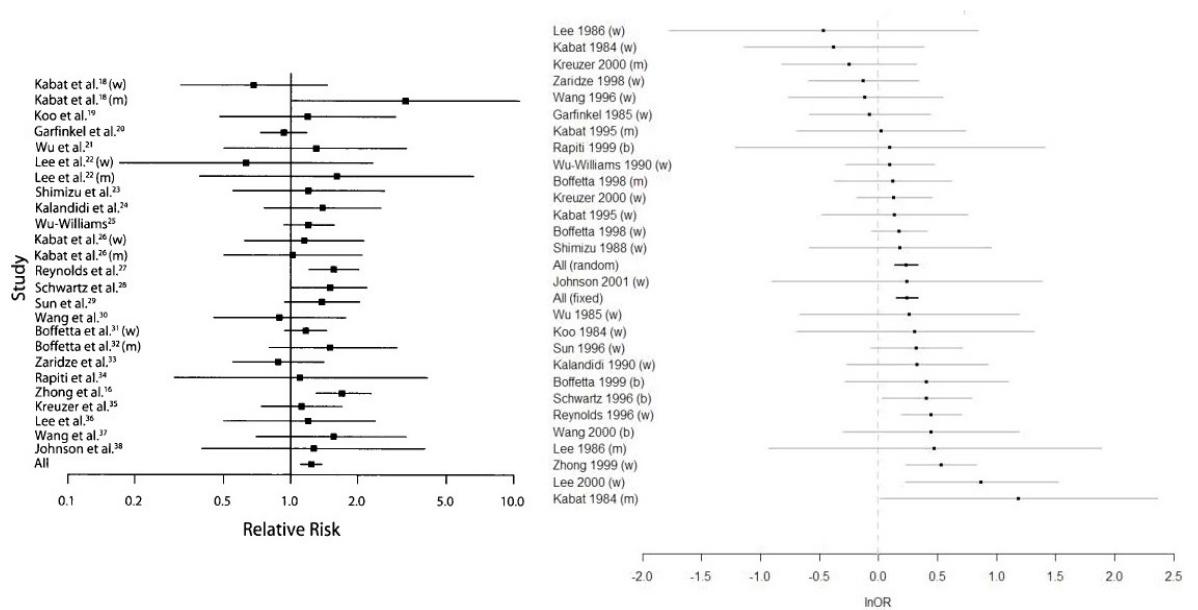


Figure 12 Comparison of the forest plot displayed in Stayner et al. (2007) on the left hand side and a version sorted by the size of the ES on the right hand side.

5.2.2 Exposure Response Analyses

In this chapter the results of the intensity and duration subgroup analyses are presented. Concerning the intensity Stayner et al. (2007) reported seven studies resulting in eight ESs aggregated to an overall highest intensity meta analysis. The result is the same for the fixed and random effects model, RR=2.01, with slightly different CIs for the fixed effects model (CI: 1.55; 2.60) and the random effects model (CI: 1.33; 2.60). The replication does not show any major differences, the fixed effects model results in an OR=2.02 (CI: 1.56; 2.60) and the random effects model in an OR=1.98 (CI: 1.50; 2.61). There is no evidence of heterogeneity ($Q = 7.59$, $df = 7$, $p = .37$), the amount of random variance is estimated $\tau^2 = .01297$.

Some differences do occur in the duration analysis replication. Stayner et al. (2007) displayed all 6 studies containing the information on duration in one of their figures (see Figure 8 for a copy of the original figure). For the regression equation a slope of $\beta = .011$ with a standard deviation of $SE = .003$ was reported, being highly significant ($p < .001$). Assuming an exposure of 45 years at work Stayner et al. predicted an RR of 1.63 (CI: 1.45; 1.82). Here the replication results in an insignificant outcome. At an $\alpha=5\%$ the calculated t-value of 1.93 is below the critical t-value of 2.12. Our analysis yields a pooled $b = .009$ and a corresponding pooled $SE = .005$. Assuming these parameters a predicted OR of 1.48 (CI: 0.96; 2.27) emerges, the CI including one. Again no sign of heterogeneity is discovered ($Q = 5.27$, $df = 5$, $p = .38$), τ^2 being close to zero. The results are displayed in Figure 13. To be able to compare all four versions, the original plot in Stayner et al. (2007), the data as it was measured and calculated by the author of this work and how the regression parameters reported at Stayner et al. look like in the replicated dataset both Figure 8 and Figure 13 are displayed in this thesis. Finally Figure 17 shows the results as they appear when using only the data as it is found in the primary studies.

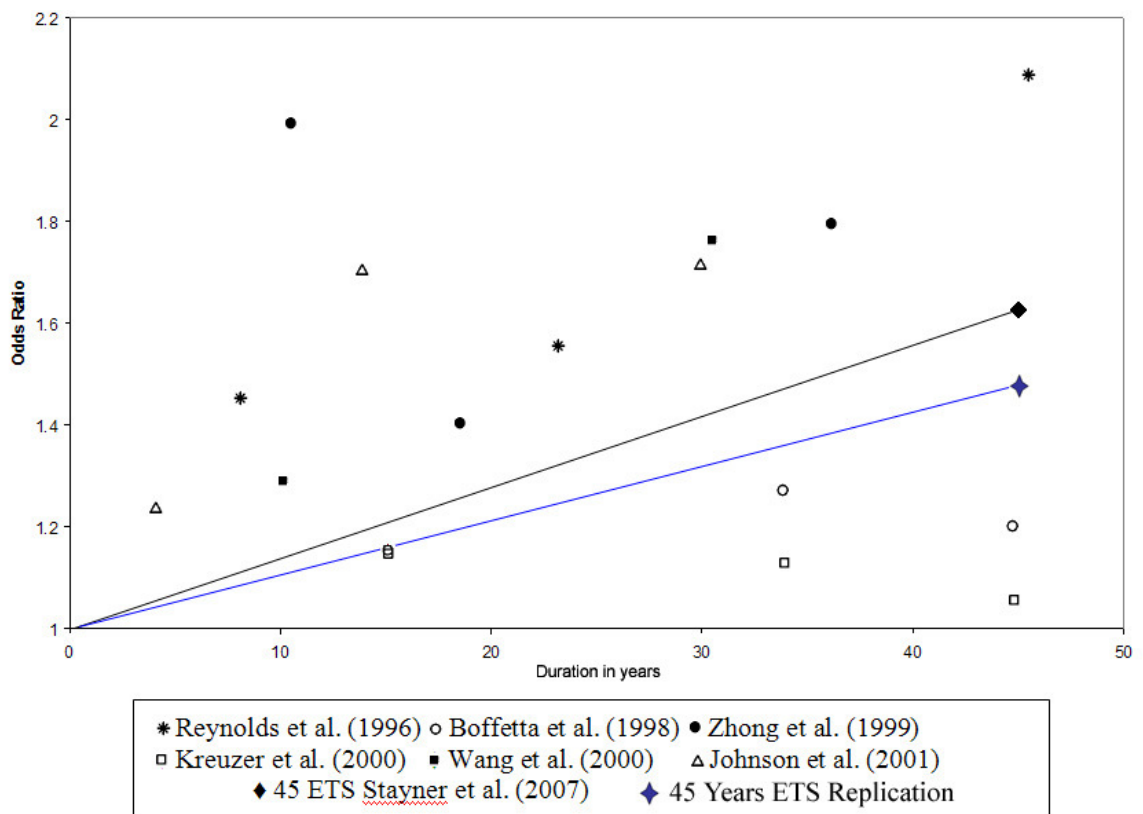


Figure 13 A plot of the replicated meta regression of the first replication phase. The upper regression line represents a line as it occurs when entering Stayner et al. (2007) regression parameters into the dataset of the first replication phase.

5.2.3 Sensitivity Analyses

Replicable sensitivity analyses were conducted in Stayner et al. (2007) for the overall mean OR, the maximum intensity analysis and the duration meta regression. Besides this a value was calculated for only the primary studies reporting adjusted ORs.

Excluding single studies from the overall effect resulted in the original study in a maximum RR of 1.27 and a minimum RR of 1.18. It is not reported whether these results are based on the fixed or random effects model. Following these analysis steps we obtain an overall OR of 1.23 (CI: 1.11; 1.35) (excluding Reynolds et al., 1996) and 1.33 (CI: 1.21; 1.46) (excluding Garfinkel et al., 1985). There is no difference using fixed or random effects model.

Confining this approach to the seven studies of the intensity analysis the original

meta analysis found RRs between 1.73 and 2.12. The replicated maximal differences are OR = 1.74 (CI: 1.29; 2.34, excluding Zhong et al., 1999) and OR = 2.24 (CI: 1.68; 2.97 excluding Kabat et al., 1984). Again there is no difference using the fixed or the random effects approach.

Using the leave-one-out-approach on the meta regression, Stayner et al. (2007) reported results varying between .009 and .014 for both β and the SE. Repeating this analysis we achieve a pooled b-weight of .006 (excluding Wang et al., 2000) and .011 (excluding either Zhong, 1999 or Kreuzer, 2000), respectively. Using a similar weight for each study a SE of .005 is calculated.

Aggregating only those studies which adjusted their reported ORs or RRs Stayner et al. (2007) presented a result of RR = 1.25 (CI: 1.13; 1.38). In the replication the same OR results for the fixed and the random effects model, an OR of 1.29 with slightly different CIs for each model, 1.18 to 1.42 and 1.16 to 1.43, respectively. No evidence of heterogeneity is found ($Q = 18.83$, $df = 16$, $p = .28$, $\tau^2 = .0070$).

5.2.4 Publication Bias

In this replication phase the publication bias analysis was replicated based on the data described in Stayner et al. (2007). The result is presented in Figure 14. On the left hand side an exact copy of the version found in the original study, on the right hand side the result of the first phase replication is displayed. R 2.6.1 and Microsoft Excel were used to plot the funnel with similar results. Here the Excel version is displayed.

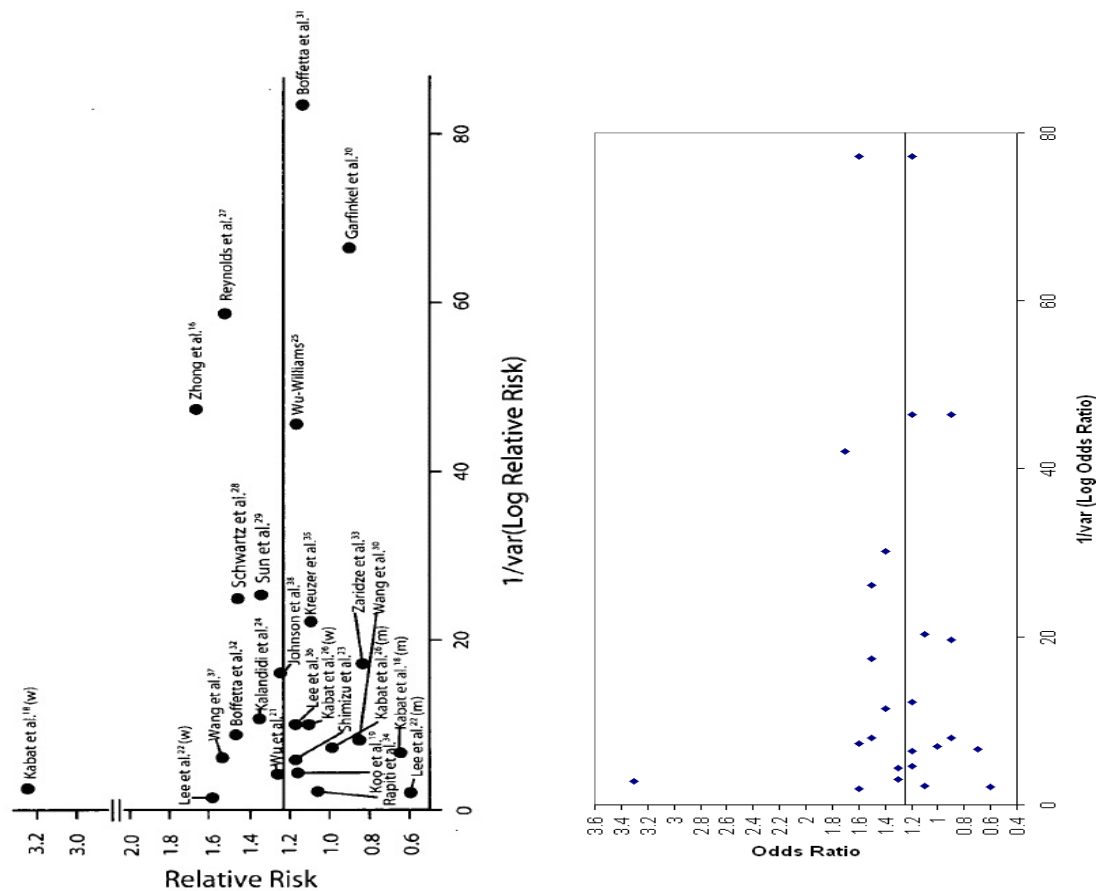


Figure 14 Comparison of the original funnel plot of Stayner et al. (2007) on the left and the exact replication on the right hand side.

Stayner et al. (2007) report, that no sign of a publication bias can be seen. At least no strong evidence for this bias can be seen in the replication as well. However there are several differences between the original and the replication. The most obvious is the break Stayner and colleagues have introduced to the RR-Axis. Kabat et al. (1985) appears much closer to the other studies, a difference which becomes quite prominent in the replication. Since one reason of a funnel plot is to show such prominent distances it is not quite clear why this break was introduced.

Taking a closer look some other differences can be discovered. Several values are shifted along the $1/\text{var}(\log \text{OR})$ -axis. The reason for these differences is not clear, how Stayner et al. (2007) arrived at their plot could not be reproduced in the present attempt.

5.3 Replication Using Data from the Primary Studies – Second Replication Phase

Here again the main aim is to follow the report structure as it was presented in Stayner et al. (2007) as close as possible. Besides the information what to replicate, no information which was reported in the original study is used, all data is derived from the primary studies. For a closer look on the data and the result outputs, see appendix B. In general the replication is rather incomplete, as several procedures were not reported in enough detail to replicate them.

5.3.1 The Overall Effect Size

Using the same selection criteria, the same 22 primary studies and the same subgroup combinations as reported in Stayner et al. (2007), Table 2 is formed. 27 ESs are discovered. Based on this data an overall mean ES of OR= 1.27 (CI: 1.16; 1.40) emerges for the fixed effects model and an OR=1.27 (CI: 1.15; 1.40) for the random effects model. The Q-test shows no evidence for heterogeneity ($Q = 27.09$, $df = 26$, $p = .40$, $\tau^2 = .0027$). Figure 15 arranges the ESs as the original study did. To provide a better comparability of the different ESs a figure which shows the ESs ordered by their size is displayed as well (see Figure 16). Here the number of significant studies is more obvious and the development of CI sizes can be followed dependent on the ES magnitude.

When using the data from the primary studies three moderators become significant. First the Boffetta group of studies ($Q = 4.09$, $p = .04$, $B = -.20$), indicating that studies related to Boffetta have a lower OR. Second and third are the adjustment for family history of lung cancer and the adjustment for occupational exposure to other carcinogens, which have to be reported together as they represent the same study sample ($Q = 4.75$, $p = .03$, $B = .29$). The adjustment for both of these results in a higher overall OR. No differences are discovered between the fixed and the mixed effects model.

Discussion

Table 2 Key Study Design Features, ORs and 95% CIs for Lung Cancer, second replication phase

Reference	Location	Time Period	Gender	N	Adjustments	Histologic Confirmation	OR	CI-	CI+
Kabat (1984)	USA	13	Men	50	Non	Yes	3.2727	1.0084	10.6212
Kabat (1984)	USA	13	Women	106	Non	Yes	.6834	.3173	1.4718
Koo (1984)	Asia	3	Women	83	Non	Yes	1.3636	.4973	3.7394
Garfinkel (1985)	USA	14	Women	329	Non	Yes	.9262	.5532	1.5504
Wu (1985)	USA	4	Women	NR	Non	Yes	1.3000	.5000	3.3000
Lee (1986)	Europe	7	Men	108	Non	No	1.6092	.3924	6.5994
Lee (1986)	Europe	7	Women	173	Non	No	.6278	.1691	2.3301
Shimizu (1988)	Asia	6	Women	NR	Non	Yes	1.2000	.6000	2.6000
Kalandidi (1990)	Europe	3	Women	207	Non	Yes	1.3889	.7594	2.5401
Wu-Williams (1990)	Asia	5	Women	1017	1, 2, 3	Yes	1.1000	.9000	1.6000
Kabat (1995)	USA	12	Men	158	Non	Yes	1.0222	.4993	2.0928
Kabat (1995)	USA	12	Women	207	Non	Yes	1.1458	.6176	2.1258
Reynolds (1996)	USA	NR	Women	876	1, 2, 3, 4, 5, 6, 7, 8, 9, 1, 11	No	1.5600	1.2100	2.0200
Schwartz (1996)	USA	12	Both	534	1, 2, 4, 9, 1, 12, 13, 14	No	1.5000	1.0000	2.2000
Sun (1996)	Asia	NR	Women	NR	1, 2	Yes	1.3800	.9400	2.0400
Wang (1996)	Asia	4	Women	270	Non	Yes	.8933	.4622	1.7263
Boffetta (1998)	Europe	10	Men	672	1, 3, 12	Yes	1.1300	.6800	1.8600
Boffetta (1998)	Europe	10	Women	1520	1, 3, 12	Yes	1.1900	.9400	1.5100
Boffetta (1999)	Europe	5	Both	247	1, 3, 12	Yes	1.5000	.8000	3.0000
Zaridze (1998)	Europe	NR	Women	547	1, 2	Yes	.8800	.5500	1.4100
Rapiti (1999)	Asia	8	Both	NR	1, 3, 12, 15	Yes	1.1000	.3000	4.1000
Zhong (1999)	Asia	7	Women	1105	1, 9, 1, 16, 17, 18, 19	Yes	1.7000	1.3000	2.3000
Kreuzer (2000)	Europe	10	Men	861	1, 3, 12	Yes	.7800	.4400	1.3800
Kreuzer (2000)	Europe	10	Women	769	1, 3, 12	Yes	1.1400	.8300	1.5700
Lee (2000)	Asia	8	Women	452	Non	Yes	2.3932	1.2509	4.5786
Wang (2000)	Asia	6	Both	NR	1, 3, 20	Yes	1.5600	.7000	3.3000
Johnson (2001)	USA	7	Women	235	2, 3, 5, 6, 21	Yes	1.2700	.4000	4.0000

1 age, 2 education, 3 location, 4 race, 5 fruit diet, 6 vegetable diet, 7 supplemental vitamins, 8 Cholesterol diet, 9 lung cancer family history, 10 occupational exposure to other carcinogens, 11 different sources of ETS, 12 gender, 13 personal history of lung related diseases, 14 usual industry, 15 religion, 16 income, 17 intake of vitamin C, 18 condition of the respondent, 19 cooking fume intensity, 20 ETS exposure at home during childhood and adulthood, 21 10-year age groups.

Discussion

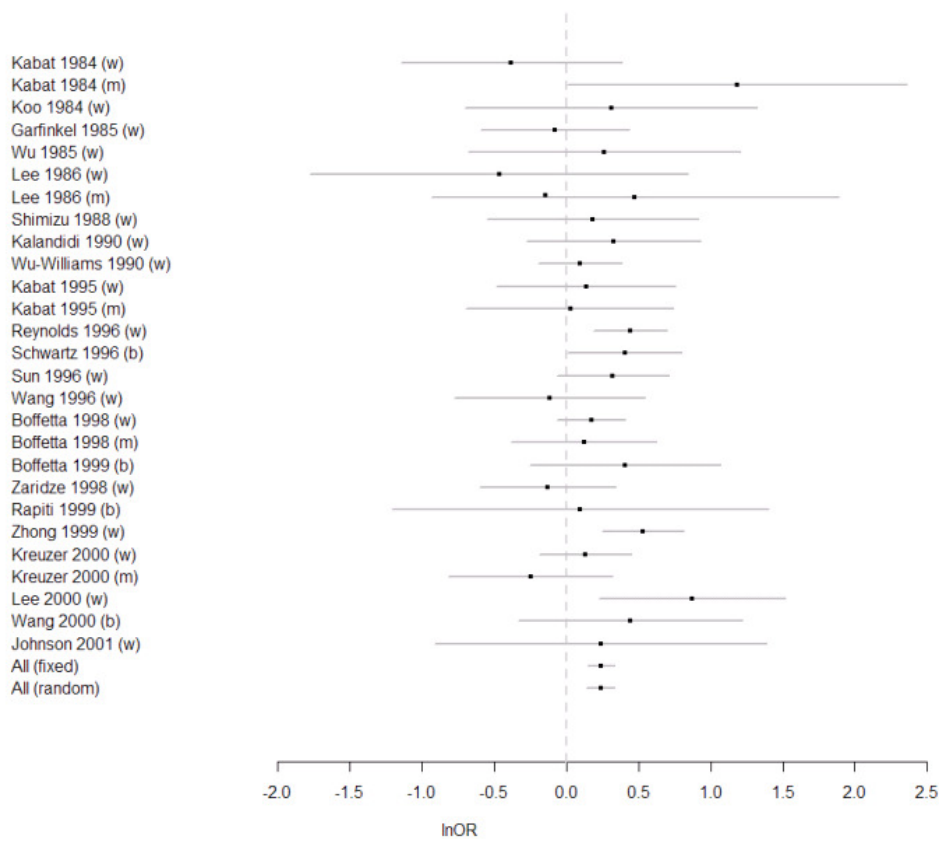


Figure 15 The forest plot based on the data coded directly from the 22 primary studies.

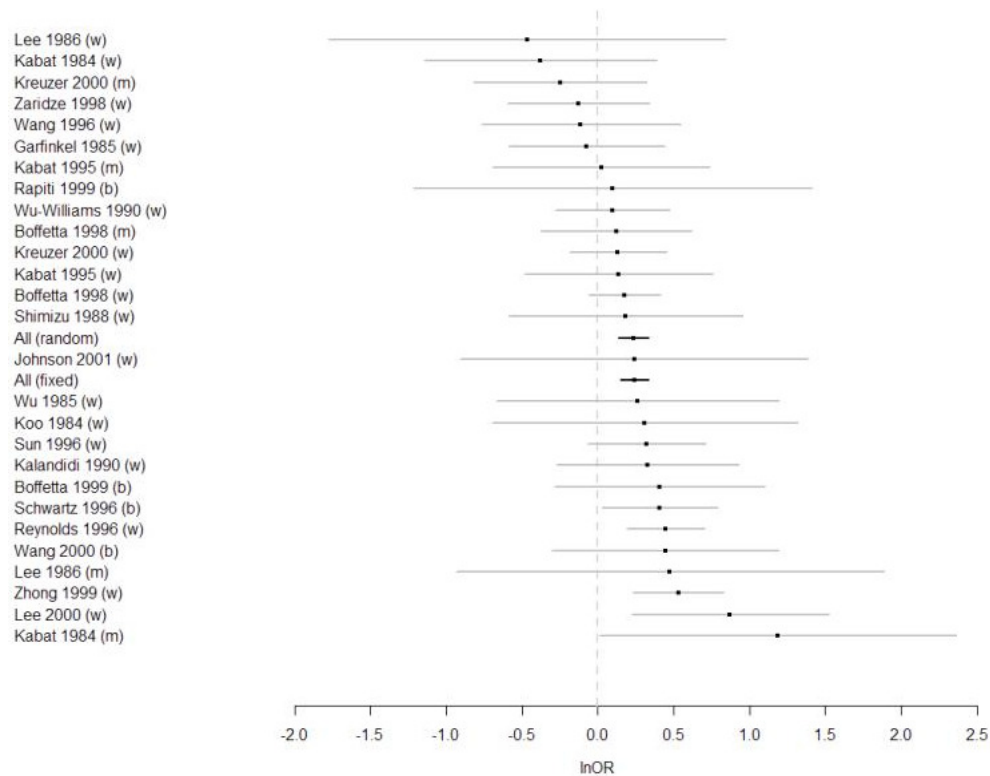


Figure 16 The same forest plot as displayed Figure 15, ordered by the ES magnitude.

Discussion

5.3.2 Exposure Response Analyses

Quite some differences appear when comparing the intensity and the duration analyses of the first and the second replication phase. Using the same guidelines as they were described in Stayner et al. (2007) more ESs to include into the intensity analysis are found. 11 ESs derived from 7 studies are included into the intensity analysis, resulting in an OR of 1.92 (CI: 1.51; 2.44) for the fixed and random effects model. Expectedly no sign of heterogeneity was discovered ($Q = 10.17$, $df = 10$, $p = .43$, $\tau^2 = .0029$). The data base for this analysis is displayed in Table 3.

Table 3 ORs and 95% CI base for the highest intensity analysis in the second replication phase

Reference	Gender	Intensity Measure	OR	CI-	CI+
Lee et al. (1986)	Men	average/a lot	.4600	.0455	4.6465
Lee et al. (1986)	Women	average/a lot	.2328	.0132	4.0952
Kalandidi et al. (1990)	Women	Some	2.1667	.8325	5.7004
Kabat et al. (1995)	Men	High	1.21	.4700	3.1300
Kabat et al. (1995)	Women	High	1.35	.6400	2.8400
Boffetta et al. (1998)	Men	More than 88.9: levels * hours/day * years	2.0000	1.0253	3.9015
Boffetta et al. (1998)	Women	More than 88.9: levels * hours/day * years	1.8700	1.1000	3.2000
Zhong et al. (Zhong)	Women	More than 4 smoking colleagues	3.0000	1.8000	4.9000
Kreuzer et al. (2000)	Men	More than 100600: hours * level of smokiness	1.0837	.3530	3.3272
Kreuzer et al. (2000)	Women	More than 100600: hours * level of smokiness	2.5180	1.1139	5.6918
Johnson et al. (2001)	Women	More than 64: Number of smoking colleagues * years at work	1.5800	.6000	4.0000

Using the meta regression approach described in chapter 4.5.2, we find a pooled b-value of .008 and an according SE of .002. Testing at an α level of .05 this regression makes a significant prediction based on the duration of ETS exposure ($t_{krit} = 2.12 < t_{ber} = 3.50$). For a 45 year exposure period this means an OR of 1.46 (CI: 1.16; 1.83). Here evidence for heterogeneity is found ($Q = 15.75$, $df = 5$, $p = .01$, $\tau^2 = .0001$), which represents another difference of the first and the second replication phase. The regression line is displayed together with the data in Figure 17.

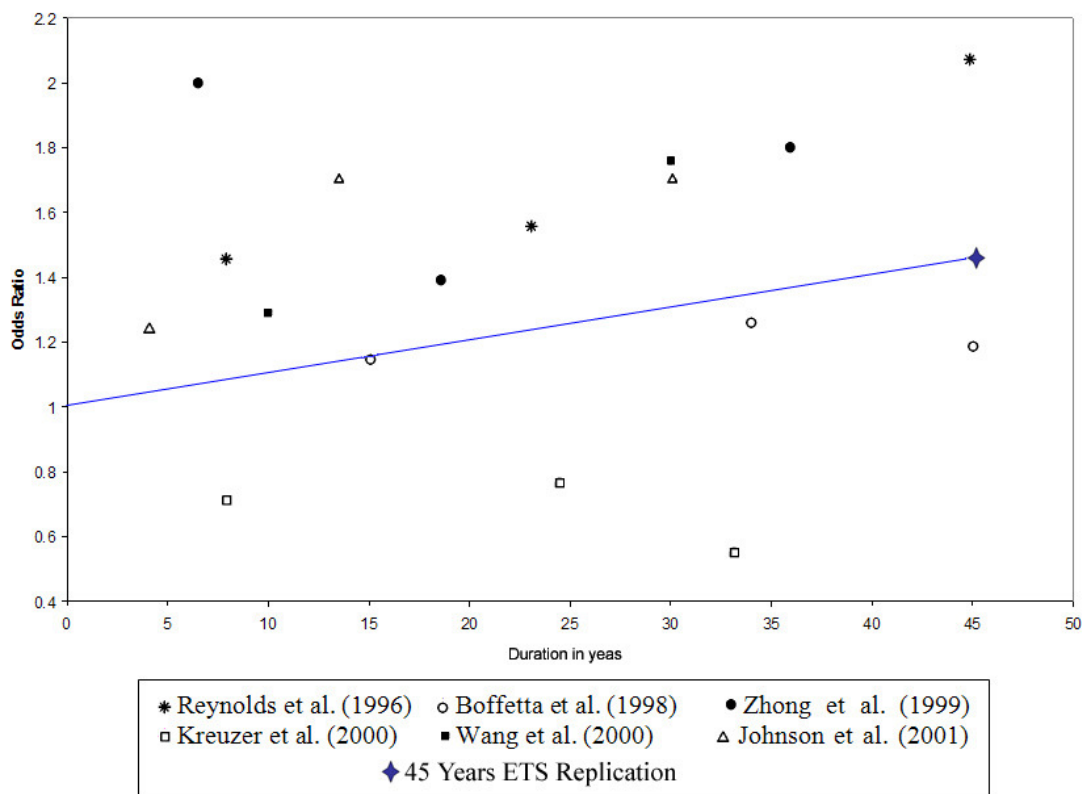


Figure 17 Replication of the meta regression predicting the OR through the duration of ETS exposure at work, second regression phase.

5.3.3 Sensitivity Analyses

Using the directives which were derived from Stayner et al. (2007) several leave-one-out analyses are presented here. A minimal value of OR = 1.23 (CI: 1.11; 1.36) arises when excluding Reynolds et al. (1996) and a maximal value of OR = 1.30 (CI: 1.18; 1.44) when excluding Kreuzer et al. (2000). Though the difference of the OR magnitude is small, compared to the first replication phase Kreuzer et al. replace Garfinkel et al. (1985), concerning the biggest influence on the maximum result.

Applying the leave-one-out approach on the intensity analysis a minimum value of OR = 1.68 (CI: 1.28; 2.20) without Zhong et al. (1999) appears. The maximum is found in 2.08 (CI: 1.60; 2.70), excluding Kabat et al. (1995). Again there is no difference using fixed or random effects models. Besides the magnitude of the ORs no differences occur between the first and the second replication phase.

The regression weights varied between b-pooled = .006 and .011, when using the

leave-one-out approach on the meta regression. For the first value Reynolds et al. (1996) is excluded, for the second Kreuzer et al. (2000). The maximum and minimum SE are .0031 and .0025, here Boffetta et al. (1998) and Wang et al. (2000) make the biggest difference. Compared to the first regression phase there is a change in the most influential primary studies, there Wang et al. 2000 stands for the smallest b-pooled value and Zhong et al. (1999) loose their place next to Kreuzer et al.

Restricting the analysis to studies which report adjusted RRs or ORs a mean OR of 1.29 is calculated, with only slightly different CIs for the fixed and the random effects model, 1.17 - 1.44 and 1.15- 1.44, respectively.

5.3.4 Publication Bias

For comparison reasons in the second replication phase the funnel plot using the inverse variance and the one using the SE are confronted (see Figure 18)

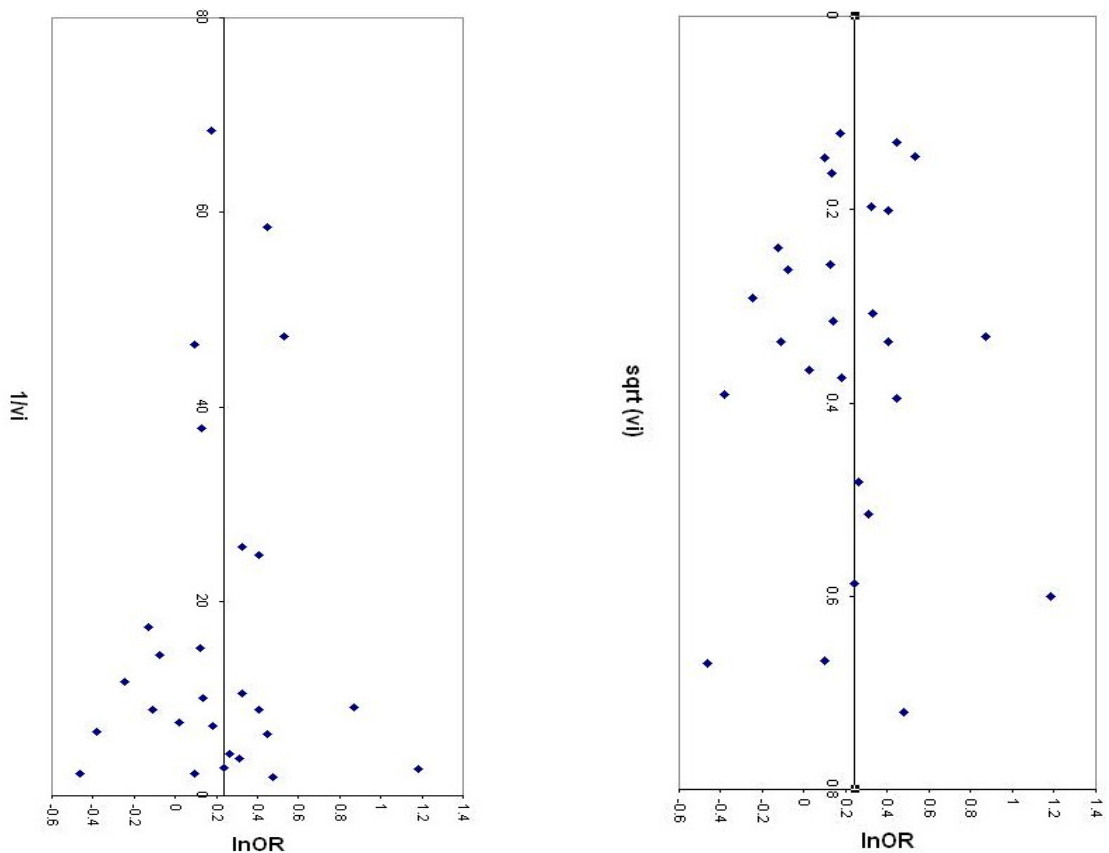


Figure 18 A comparison of a funnel plot using the inverse variance (left hand side) and one that is using the SE (right hand side), both based the data of the second replication phase.

Though there are differences between the two plots no clear evidence of a publication bias can be seen in both funnels.

5.4 Results of the Crossroad Model

Following the described approaches of chapter 4.6 here the influence magnitude of the CRM is shown. The “bad guy” analysis uses the smallest RRs or ORs of the 22 available primary studies which were discovered by the author, changing the way through the CRM and the present sub groups at free will. For the fixed effects model an overall OR = 1.04 appears, with a CI of 0.94 to 1.15. For the random effects model we found an OR = 1.03 with a CI of 0.92 to 1.15. No sign of heterogeneity is evident ($Q = 22.88$, $df = 21$, $p = 0.35$). The results of the good guy analysis can be found in Pahl (2009).

Analysing the four theories described in chapter 4.6.2 a quite differentiated picture appears. Judged by ES magnitude the third theory produces the strongest result. Aggregating the four studies reporting the rigorous ETS group definition an OR of 1.45 (CI: 1.05; 1.99) appears. The fourth theory, the lax ETS group definition, provides the next bigger mean ES with an OR = 1.40 (CI: 1.23; 1.61) for the fixed effects model and an OR = 1.44 (CI: 1.14; 1.82) for the random effects model, based on nine ESs from primary studies. Third in row is theory one that includes only women with the second ETS group definition and unadjusted ORs, resulting in a mean OR of 1.33 (CI: 1.20; 1.47) for the fixed effects model and an OR of 1.28 (CI: 1.09; 1.51) for the random effects model, both based on 14 ESs. This value is followed by the same theory which uses only adjusted values. This theory results in a mean OR of 1.23 (CI: 1.08; 1.40), regardless of the underlying model. The second theory thus produces the smallest mean ORs with a fixed effects model mean of OR=1.11 (CI: 0.83; 1.48) for unadjusted results concerning men and a random effects model mean of OR =1.13 (CI: 0.79; 1.61). Finally the adjusted mean OR for theory two is OR = 0.87 (CI: 0.66; 1.14) for the fixed and OR= 0.89 (CI: 0.63; 1.27) for the random effects model, respectively. All results of the second theory are based on only two studies, Boffetta et al. (1998) and Kreuzer et al. (2000).

Discussion

All resulting overall ESs of the CRM are displayed in Figure 19. To provide a better comparability of the theories they are sorted by ES magnitude. Additionally the number of studies which are included into the aggregated OR are shown (N). All data sets and result outputs can be found on the appendix CD in section C.

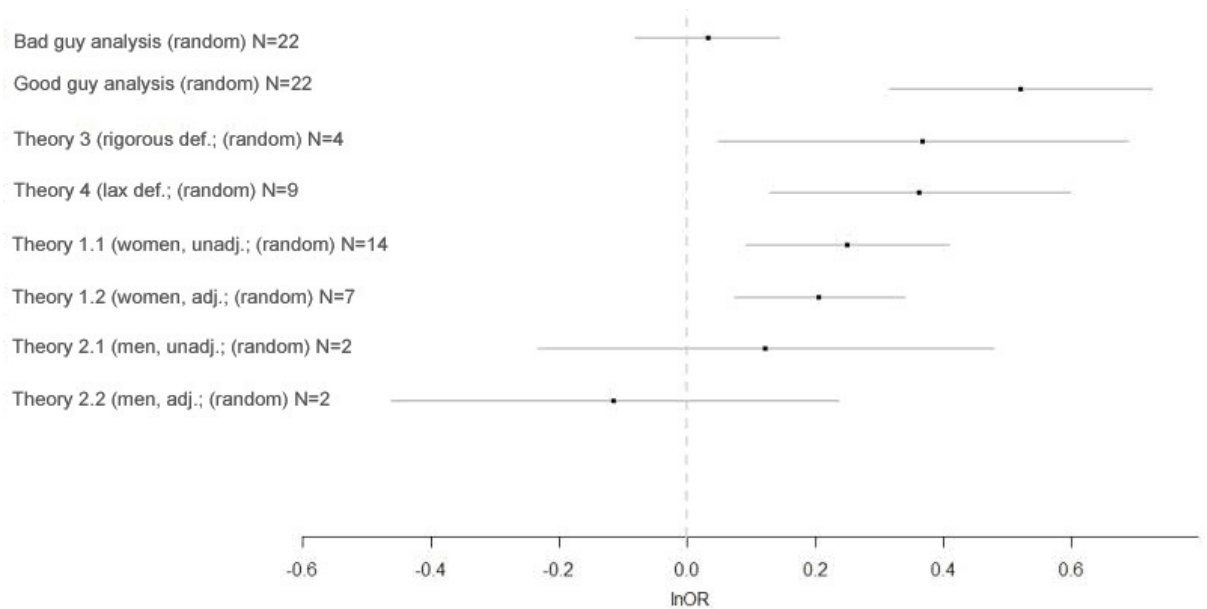


Figure 19 Forest Plot displaying the good any bad guy analysis as well as the four theories for the Crossroad Model.

5.5 Advanced Analyses

In this chapter the results for the additional analysis proposed for this study question are presented. The datasets for these analyses and the result outputs are found on the appendix CD section D.

5.5.1 Quality of Primary Studies

A quality rating based on the fourth coding block is used as a moderator in this analysis. Of the 10 possible points a range of 8 (Lee et al., 2000) and 1 (Reynolds et al., 1996) points are assigned. This moderator is used on the data provided directly in Stayner et al. (2007) and on the data of the newly coded primary studies. On the 25 ESs presented in Stayner et al. the moderator has a significant influence ($b = -.0472$, $Q = 4.3431$, $p = .0372$), indicating a smaller OR with an increasing quality. The same quality moderator does not lead to significant results when applied to the

27 ESs of the second replication phase ($b = -.0424$, $Q = 3.09$, $p = .08$), though the general direction of the effect is the same. All results are based on logarithmized ORs. For both replication phases appear no differences between the fixed and the mixed effects model.

5.5.2 Correction of Study Artefacts

In chapter 4.7.2 the need for constant reliability parameters is described. Ashton (2000) reported a mean test-retest reliability coefficient of $r_{yy} = .73$ for medical and psychological doctors. This value is used to correct the dependent variable side of this thesis, the judgment whether a patient has lung cancer or not. A kappa of $r_{xx} = .46$ is taken from Pron et al (1988) which stands for the reliability of patient's reports on exposure to occupational passive smoke. Using equation 22 a disattenuation factor $a = .58$ is calculated. Table 4 shows the 22 uncorrected ES parameters applicable for this correction and an according mean OR in comparison to the corrected parameters. Due to the excluded studies an uncorrected mean OR of 1.24 (CI: 1.09; 1.40) is calculated. This value has to be compared to the corrected mean OR of 1.40 (CI: 1.13; 1.73). Thus the overall OR is increased by the correction procedure, alongside with a widened CI. Homogeneity analysis results in an explained variance of 16.7%, which is considerably lower than the 75% rule of thumb introduced by Hunter & Schmidt (2004).

Discussion

Table 4 Results of the correction of reliability on both the independent and the dependent variable side. On the left side the original parameters are displayed, marked with ORO and CIO. On the right side the corrected parameters are marked with ORC and CIC

Studie	ORo*	CIO**	ORc	CIC
Kabat 1984 (w)	0.68	0.53; 0.87	0.51	0.32; 0.79
Kabat 1984 (m)	3.27	2.11; 5.45	11.47	3.92; 1.0151***
Koo 1984 (w)	1.36	1.30; 1.43	1.72	1.57; 1.87
Garfinkel 1985 (w)	0.93	0.83; 1.04	0.88	0.72; 1.06
Lee 1986 (w)	0.63	0.47; 0.83	0.44	0.25; 0.72
Lee 1986 (m)	1.61	1.43; 1.82	2.31	1.86; 2.91
Kalandidi 1990 (w)	1.39	1.31; 1.47	1.77	1.60; 1.96
Wu-Williams 1990 (w)	1.10	1.06; 1.15	1.18	1.10; 1.26
Kabat 1995 (w)	1.15	1.12; 1.17	1.27	1.21; 1.32
Kabat 1995 (m)	1.02	0.95; 1.10	1.04	0.92; 1.17
Reynolds 1996 (w)	1.56	1.40; 1.74	2.19	1.81; 2.67
Schwartz 1996 (b)	1.50	1.37; 1.64	2.04	1.73; 2.40
Wang 1996 (w)	0.89	0.79; 1.02	0.82	0.66; 1.03
Boffetta 1998 (w)	1.19	1.18; 1.20	1.35	1.33; 1.37
Boffetta 1998 (m)	1.13	1.10; 1.16	1.24	1.17; 1.30
Boffetta 1999 (b)	1.50	1.37; 1.64	2.04	1.73; 2.40
Zaridze 1998 (w)	0.88	0.77; 1.01	0.80	0.63; 1.01
Zhong 1999 (w)	1.70	1.47; 1.97	2.56	1.97; 3.39
Kreuzer 2000 (w)	1.14	1.11; 1.17	1.25	1.20; 1.31
Kreuzer 2000 (m)	0.78	0.65; 0.94	0.65	0.47; 0.90
Lee 2000 (w)	2.39	1.79; 3.28	5.12	2.81; 11.54
Johnson 2001 (w)	1.27	1.25; 1.29	1.51	1.46; 1.57
All (random)	1.24	1.09; 1.40	1.40	1.13; 1.73

* These are the original values which are calculated to correlations and back for comparison reasons, ** these CIs are calculated by the reported study N where ever possible, values will be different compared to the Stayner et al. (2007) values, *** this is a correlation. After the correction this value tunes out bigger than one and can thereby not be formed back to an OR.

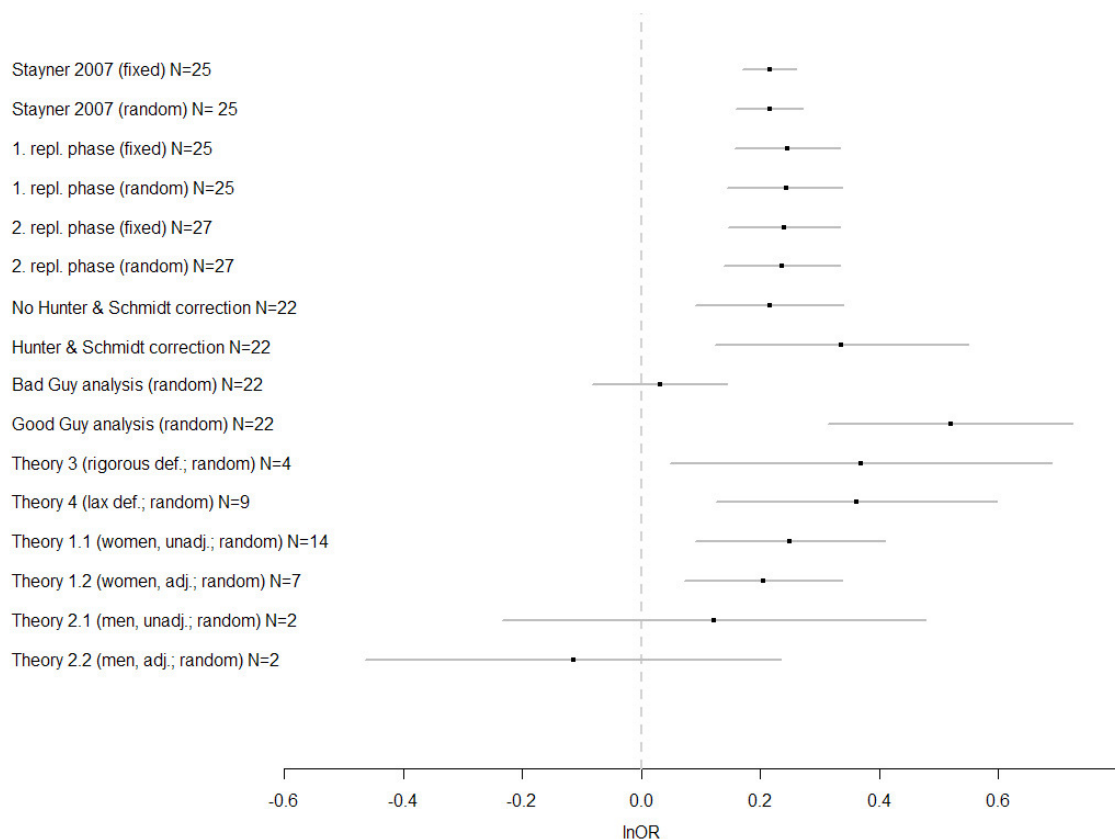
5.6 Summary

To enhance the comparability of the outcomes in this chapter all results are summarized. To be able to judge the robustness of the results of Stayner et al. (2007) their results have to be compared to the two replication phases. Table 5 recapitulates the results described above. These results have to be compared to the bad and good guy analyses and the four theories of the crossroad model to show the influence of subjectivity on the study question at hand. To accomplish that a forest plot is displayed in Figure 20. Here all mean ESs related to this thesis are plotted against each other, ES to be compared are plotted closer to each other.

Discussion

Table 5 Comparison of the original study and the two replication phases

	Results Stayner et al. (2007)	Results first replication phase	Results second replication phase
Overall ES	N= 25 FE: RR=1.24 (1.18; 1.29) RE: RR=1.24 (1.17; 1.31)	N= 25 FE: OR=1.28 (1.17; 1.40) RE: OR=1.27 (1.16; 1.40)	N= 27 FE: OR=1.27 (1.16; 1.40) RE: OR=1.27 (1.15; 1.40)
Intensity analysis	N= 8 FE: 2.01 (1.55; 2.60) RE: 2.01 (1.33; 2.60)	N= 8 FE: 2.02 (1.56; 2.60) RE: 1.98 (1.50; 2.61)	N= 11 FE: OR=1.92 (1.52; 2.44) RE: OR=1.92 (1.51; 2.44)
Meta regression	β =.011; SE=.0025	b=.0087; SE=.0045	b=.0084; SE=.0024
Predicted OR 45 years ETS exposure	RR=1.63 (1.45; 1.82)	OR=1.48 (.96; 2.27)	OR=1.46 (1.16; 1.83)
Leave-one-out: all 22 studies	RR=1.18-1.27	OR=1.23-1.33	OR=1.23-1.30
Leave-one-out: Intensity analysis	RR=1.73-2.12	OR=1.74-2.24	OR=1.68-2.08
Leave-one-out: meta regression	β =.009-.014 SE=.009-.014	b=.0058-.0110 SE=.005	b=.0055-.0114 SE=.0025-.0031
Overall ES, adjusted values only	N= 17 RR=1.25 (1.13-1.38)	N= 17 FE: OR=1.29 (1.18; 1.42) RE: OR=1.29 (1.16; 1.43)	N= 14 FE: OR=1.29 (1.17; 1.44) RE: OR=1.29 (1.15; 1.44)

**Figure 20 Forest plot of all mean ESs. Differences of the replication process and the crossroad model can be compared.**

6 Discussion

Subjectivity and its effect on results of a meta analysis is the central subject of the thesis at hand. Which decisions have to be made during the conduction of a meta analysis? What are the differences between different replications? Which crossroads have a specific effect on the results of ETS at work and lung cancer? How robust are the results in this study field? For this reason the meta analysis of Stayner et al. (2007) was replicated by firstly using the data provided in the original study and secondly by collecting the same primary studies and repeating the same analysis steps. Then a sample of possible crossroads was chosen to directly analyse the effect of decisions. A theory free “good guy-bad guy” approach was used to show the maximum impact of all crossroads combined. Four more realistic approaches were chosen to show the effect of decisions in a real world setting. Finally a set of additional analysis methods were applied to the study sample to increase the knowledge gain.

6.1 Interpretation of the Results

(I) Are the results given in Stayner et al. (2007) to be replicated?

Comparing the results of the original meta analysis and the results of the first replication phase no significant differences appear. However, this is only true for the analyses which could have been replicated, due to poor documentation several procedures were not tested at all. The main difference appeared comparing the CIs. The replicated CIs are wider than the original ones. It is not clear where this difference comes from. It is assumed that different, not reported methods for CI calculations were used in Stayner et al. (2007) and that the statistic programs used might have a certain impact.

(II) Is the replication of the results of Stayner et al. (2007) possible if the data is to be newly coded?

This question can partially be answered positively. Overall results do not vary much, however the data basis shows several differences. Even though the same procedure guidelines were followed as they are reported in Stayner et al. (2007) the

author of this thesis and Pahl 2009 found 27 ESs instead of the 25 in the original study. This might be caused by implicit additional rules not reported in the original study. The same picture appeared when replicating the intensity subgroups analysis, here eleven ESs were found instead of eight. This difference only had a minor effect on the overall result; the mean ES is quite similar.

The biggest differences were discovered in the meta regression. Slope and constant were smaller, the test for heterogeneity became significant and the replicated figure communicates a quite different impression compared to the original one. This difference was difficult to interpret, as it can have several sources. First Kreuzer et al. (2000) might have had a major impact, as there were cases present in the second replication which were not included in the original analysis. This difference has effect to some extent on each of the presented analyses, with the exception of the moderator analysis concerning the intensity of ETS exposure. Second the weighting process was poorly documented; there might be differences in the procedures. Finally the rule to use the middle of the duration period was not followed by Stayner et al. (2007) in all cases (for Kreuzer et al., (2000) apparently the highest limit was used), this will cause some additional differences.

It was not possible to replicate the funnel plot in either the first or the second replication phase. This might be caused partially by rounding differences. These normally minor changes become a much bigger emphasis through the nature of the inverse variance, as the effect is multiplied by using this approach. Another reason might be the use of different equations to indirectly calculate the variance. Stayner et al. (2007) did not report which of the three possibilities they used (see chapter 4.4.1, equations 4 and 5). Again this effect is multiplied by using the inverse variance. A common recommendation to avoid this multiplication of small differences is the use of the standard error instead of the inverse variance (Sterne & Egger, 2001).

(IIIa) Do the proposed crossroads have a significant impact on the results of the present meta analysis?

The results of this thesis clearly showed an effect of the selected crossroads.

Depending on the position of points different numbers of ESs were applicable to the analysis and the results showed a wide range.

(IIIb) What is the minimum RR contained in the data and which way leads to this result?

The bad guy result is significantly below the original mean ES of 1.24 (1.17; 1.31), with a result of OR = 1.03 (.92; 1.15). The good guy result is significantly above the original mean OR.

(IIIc) What is the span between the chosen theories?

Using the approach of theories stated in advance and applying them on the present data set led to another set of different results. Here the CIs did overlap, however this is mainly due to the comparatively small sample sizes. At least this leads to the conclusion that more research is needed addressing the subgroups which were underrepresented in this study sample. Nevertheless the difference between the gender groups and the adjustment groups seem promising for further investigation in a larger study sample.

(IV) Does the quality rating of the primary studies explain a significant part of the data variation?

Quality ratings of block four had a significant impact on the study results on the original dataset. Taking into account that most of the validity questions of block could not be answered it was surprising that the rest still held predictive value. This is another hint for the advantage of a high level documentation. In this way conductors of systematic reviews would be able to correct for possible quality problems to an even bigger extend.

(V) Does the introduction of study artefacts shed additional light on the research question of ETS at the work place?

The introduction of the Hunter and Schmidt approach (2004) revealed on the one hand that documentation quality and detail needs to be immensely increased to enable most of the correction procedures. The two corrections conducted showed an

increase of the mean OR and the CI range. This difference was not significant; the CI ranges did not overlap. Still the effect two of at least twelve possible corrections had seems promising enough to increase the effort to include further artefact corrections into health matter analyses. Especially as the explained variance of 16.7% was considerably lower than the 75% rule of thumb introduced by Hunter and Schmidt (2004).

6.2 Limitations

Judging the literature search process of Stayner et al. (2007) it is not likely that a complete study sample was accomplished. Additionally the search took place in 2001, the article was published in 2007. It is possible that in the meanwhile new studies were published on the topic of ETS at work and lung cancer. An additional literature search would have solved these questions, however the author of the meta analysis at hand and Pahl (2009) decided against this procedure. Here we concentrated on methodological differences of studies with the same data foundation. Adding more and newer studies to the sample would have enhanced the quality of the conclusion on ETS and lung cancer but would have distracted from the comparison approach introduced above. Still the insertion of more studies might help future investigations about the CRM to increase the number of crossings analysed.

During the coding process it became apparent that many items designed for the study question could not be answered by most studies of the sample. In block four more than 50 items had to be excluded due to the lack of reported information, in the other blocks missing data was smaller but present as well. Taking the additional knowledge gained by those few analysable items of this thesis into account the benefit of a high documentation standard might be very impressive.

Stayner et al. (2007) described a sensitivity analysis, testing the difference of the original meta analysis and Wells (1998), who was supposed to have used the same studies, with only some newer publications missing. However, this process was reported incompletely in the original study, and analyzing Wells (1998) did not

clarify the procedure as well. Thus the replication of this analysis step could not be done.

During the replication process the problem appeared that the method of weighting in the duration regression was not reported at all in Stayner et al. (2007). In the first replication no weights were used and in the second replication phase weights of our own judgment were applied. Differences between the original and the replication results might be partly caused by this problem. The same is true for the correction of dependent data points of the replication. One of the basic requirements for a regression is the independence of the data points entering the regression (Bortz, 2005). If there are several values from one study entering a regression correction methods for their interdependence must be applied. For the first regression phase not enough information was reported in the original study. For the second replication Wang et al. (2000) did not meet the requirements the correction method described by Greenland and Longnecker (1992). Stayner et al. did not report how they solved this problem (no case or control numbers), thereby the approach to replace the missing value with the mean of the other primary studies might be a different one than in the original study. In this meta analysis the missing value was replaced by the mean which resulted calculating the parameters of the other studies.

As stated in chapter 4.4.1 SPSS 14.0 was mainly used for the replication, in addition to R 2.6.1, MetaWin 2.1 and Microsoft Excel 2003. Some difficulties with this approach might be found as not all statistic analysis software uses the same procedures. There is some evidence that Stayner et al. (2007) might have used SAS (SAS Institute Inc, Cary, NC) for all their analysis steps, as they have used it for their meta regression. However, SAS employs the t-distribution in order to determine the CI. By default the degrees of freedom should have been $df = k - 1$, this default can be changed, though. Some of the variation between Stayner et al. (2007) results and the results of this thesis might be based on this difference. The choice of the statistical analysis package might be used as another crossroad, yet this approach was not followed more closely in this thesis.

One critique passed on Stayner et al. (2007) is partially true for this thesis as well. In the first and second replication the same mixture of adjusted and unadjusted ESs were added into the analysis. In the CRM the third and fourth theory made no absolute declaration on whether adjusted or unadjusted ORs should be added into the aggregate. With a larger number of available ESs a moderator analysis on the effects of different adjustments would have been possible. In theory one and two the mixture of adjusted and unadjusted ESs was avoided, which decreased the available number of ESs significantly. For theory 2.1 and 2.2 only two eligible studies were left, Boffetta et al. (1998) and Kreuzer et al. (2000). This is not only a very small number of ESs but also a problematic result as apparently a number of cases and controls were the same in both studies. Interpretation of the bad guy analysis is difficult as well, as here adjusted, unadjusted and subgroup results were aggregated.

There are measures in this thesis which are prone to subjectivity as well. Especially the quality ratings of block two and four and the rating of the 35 MOOSE items will be influenced by some kind of subjectivity, but even if some points are changed to a more positive outcome, the overall methodological quality of Stayner et al. (2007) will still seem rather poor. To be able to expose the amount of subjectivity in the quality rating a higher number of coders would have helped quantifying the subjectivity. On the other hand the interrater reliability for three coders was sufficient, the amount of subjective variation should not be overwhelming.

Only a small number of crossroads could have been analysed in this thesis. The three chosen subjective decisions were part of the data generation phase of a meta analysis. In some ways they may resemble moderators, however there are elemental differences. Moderators are captured study properties which are assumed to have influence on an independently assembled data base. On the other hand crossroad decisions in the data generation phase limit the eligibility of certain data points to enter the analysis in the first place. The goal of the CRM in the thesis at hand was to expose these subjective decisions as they appeared in Stayner et al. (2007). After discovering possible subjective decisions possible specifications of those crossroads

were defined, which were added to the coding sheet. Other considered but not elaborated crossroads do not have this similarity to moderators, e.g. the choice of the statistic program or of the statistical procedures. Both examples belong to the evaluation of data points. Other authors have addressed these crossroads under a different name, so called “judgement calls” (e.g. Wanous, 1989). For a more complete picture of the effects of crossroads a full-fetched model would be preferable.

The Hunter and Schmidt (2004) study artefact correction was a rather exploratory attempt to increase the explained variance. The authors of the meta analysis at hand did not find any other ETS related analyses which tried this approach before. Since the conversion of ORs into correlations and back in connection with the Hunter and Schmidt approach was not tested before, further investigation is needed to make sure this approach is applicable. In the end only two artefacts could have been corrected, the reliabilities of the dependent and the independent variable. To accomplish this, constant values for each variable had to be used. However, a direct correction of each ES depending on the quality of their measurement would have been more appropriate. The two constant reliability estimates derived from Ashton (2000) and Pron et al. (1988) might be misleading, other authors have different results (e.g. Huang 2008). A replication of our results with different overall reliability estimates or, even better, with suiting reliability estimates for each primary study should be attempted. Incidentally this opens an opportunity to add another crossroad: Which artefact estimates are assumed in an analysis?

Additionally different artefacts might arise in a study concerning health matters compared to the industrial and organizational inspired approach of Hunter and Schmidt. These new artefacts might suit the available information better than the original twelve artefacts did. Overall, particularly an increase documentation density would help adding more artefact corrections to health matter analyses.

6.3 Conclusion

In this thesis the crossroad model was introduced and implemented on a study of

ETS at work and lung cancer. The effect of subjectivity was shown and a range of possible meta analysis results due to subjectivity was discovered. A replication by an independent research group did help to increase the standard of knowledge of the topic at hand and discovered further open questions. However, only a small part of possible crossroads and advanced analysis methods were used in the present meta analysis. More attention needs to be spend on the other steps of meta analyses, especially on the data analysis and the presentation of results. Hence, a foundation was laid for further investigations, where new, health related artefacts and crossroads might be found which significantly influence study results. Thereby a clearer picture of health topics might emerge. In order to accomplish this, documentation quality should be increased in primary studies and in meta analyses. All procedural steps should be reproducible; all decisions and compromises made concerning the study question should be published. Only then a reader of a meta analysis can judge whether the way travelled through the according CRM suits his or her own opinion. Thereby one of the major drawbacks of meta analysis will be solved, by shared subjectivity (DeCoster, 2004).

Additionally, interdisciplinary communication would enhance the advancement in science significantly. New method developments in each field can trigger new results in other fields; new conclusions in one area might animate new ideas in another. Hopefully the combination of psychological methods with a health subject was one step into this direction. After all the methodological sophistication of meta analyses can still be increased (Sutton & Higgins, 2008).

7 Reference

Primary studies are marked with (*).

- Abrami, P., C., Cohen, P., A. & d'Appolonia, S. (1988). Implementation Problems in Meta-Analysis. *Review of Educational Research*, 58(2), 151-179.
- Ashton, R., H. (2000). A Review and Analysis of Research on the Test-Retest Reliability of Professional Judgment. *J. Behav. Dec. Making*, 13, p 277-294.
- Barnes, D., E. & Bero, L., A. (1998). Why review articles on the health effects of passive smoking reach different conclusions. *Journal of the American Medical Association*, 279, 1566-1570.
- Beecher, H., K. (1955). The powerful placebo. *The Journal of the American Medical Association*, 159, 1602-1606.
- Biggerstaff, B. J., Tweedie, R. L., Mengersen, K. L. (1994). Passive smoking in the workplace: Classical and Bayesian meta-analyses. *International Archives of Occupational and Environmental Health*, 66, 269-277.
- *Boffetta, P., Agudo, A., Ahrens, W., Benhamou, E., Benhamou, S., Darby, S. C., et al. (1998). Multicenter case-control study of exposure to environmental tobacco smoke and lung cancer in Europe. *Journal of the National Cancer Institute*, 90, 1440-1450.
- *Boffetta, P., Ahrens, W., Nyberg, F., Mukeria, A., Brüske-Hohlfeld, I., Fortes, C., et al. (1999). Exposure to environmental tobacco smoke and risk of adenocarcinoma of the lung. *International Journal of Cancer*, 83, 635-639.
- Booth, A. (2006). "Brimful of STARLITE": Toward standards for reporting literature searches. *Journal of the Medical Library Association*, 94, 421-429.
- Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler*. 4th edition. Berlin: Springer.
- Bortz, J. (2005). *Statistik für Human- und Sozialwissenschaftler*. 6th edition, Heidelberg: Springer Medizin Verlag.
- Bosnjak, M. & Viechtbauer, W. (2009). Die Methode der Meta-Analyse zur Evidenzbasierung von Gesundheitsrisiken: Beiträge der Sozial-, Verhaltens- und Wirtschaftswissenschaften. *Zbl. Arbeitsmed*, 59.

Reference

- Bosnjak, M., Viechtbauer, W. (in press). Die Methode der Meta-Analyse zur Evidenzbasierung von Gesundheitsrisiken: Beiträge der Sozial-, Verhaltens- und Wirtschaftswissenschaften.
- Bradbury-Jones, C. (2007). Enhancing rigor in qualitative health research Exploring subjectivity through Pershkin's I's. *Journal of Advanced Nursing*, 59(3), 290–298.
- Brüderl, J. (2004). Meta-Analyse in der Soziologie: Bilanz der deutschen Scheidungsursachenforschung oder "statistischer Fruchtsalat"? *Zeitschrift für Soziologie*, 33, 84–86.
- Brügelmann, H. & Heymann, H., W. (2006). Klärung und Übersetzung von Forschung als Dienstleistung für die pädagogische Praxis. Retrieved October, 21, 2008 from <http://www.agprim.uni-siegen.de/printbrue/bruehwhevaforsch.pdf>.
- Chappell, W., R. & Gratt, L., B. (1996). A graphical method for pooling epidemiological studies. *American Journal of Public Health*, 86, 748–750.
- Choi, J. (2006). Doing poststructural ethnography in the life history of dropouts in South Korea: methodological ruminations on subjectivity, positionality and reflexivity. *International Journal of Qualitative Studies in Education*. 19(4), 435–453.
- Cohen, J., Cohen, P., West, S.,G. & Aiken, L., S. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. 3rd Ed. Lawrence Erlbaum Associates.
- Cook, T., D. & Campbell, D., T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin Company.
- Cooper, H., M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, 52, 291–302.
- Cooper, H. & Hedges, L., V. (1994). Research synthesis as a scientific enterprise. In H. Cooper, & L., V. Hedges (Ed.), *The Handbook of Research Synthesis*, New York, NY: Russel Sage, 3–14.
- Cooper, H., Hedges, L., V. & Valentine, J., C. (2008). *The handbook of research synthesis* (2nd Ed.). New York, NY: Russell Sage Foundation.

Reference

- Copas, J. (1999). What Works?: Selectivity Models and Meta-analysis. *J.R. Statist. Soc. A*, *162*(1), 95-109.
- Covey, J. (2007). A Meta-analysis of the Effects of Presenting Treatment Benefits in Different Formats. *Medical Decision Making*, *27*, 638-654
- DeCoster, J. (2004). *Meta-analysis Notes*. Retrieved October, 10, 2008 from <http://www.stat-help.com/notes.html>.
- Deeks, J., J. (1998). When can odds ratios mislead? *British Medical Journal*, *317*, 1155-1156.
- DerSimonian, R. & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*, 177-188.
- Deutscher Bundestag, (2007). *Drucksache* 16/5049.
- Doll, R. & Hill, A., B. (1950). Smoking and carcinoma of the lung. *British Medical Journal*, *2*, 740-748.
- Drinkmann, A. (1990). *Methodenkritische Untersuchungen zur Meta-Analyse*. Weinheim: Deutscher Studien Verlag.
- Duval, S., J. & Tweedie, R., L. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*, 455-463.
- Egger, M. & Davey S., G. (1997). Meta-Analysis: Potentials and promise. *British Medical Journal*, *315*, 1371-1374.
- Egger, M., Davey Smith, G., Schneider, M. & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629-634.
- Endrass, J., Rosegger, A. & Urbaniok, F. (2007). *Züricher Forensik Studie Abschlussbericht des Modellversuchs: Therapieevaluation und Prädiktorenforschung*. Retrieved November, 5, 2008 from http://www.zurichforensic.org/1/1_2/PDF_Dokumente/ABSCHLUSSBERICHT_MV_November2007_defintiv.pdf
- Eysenck, H., J. (1978). An exercise in mega-silliness. *American Psychologist*, *33*, 517.

Reference

- Field, A., P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics*, 2, 105-124.
- Franke, G., R. (2001). Applications of Meta-Analysis for Marketing and Public Policy: A Review. *Journal of Public Policy & Marketing*. 20(2), 186-200.
- Frantz, J., (2000). *G3data 1.5.1*. From <http://www.frantz.fi/software/g3data.php>.
- Fricke, R., & Treinies, G. (1985). *Einführung in die Metaanalyse*. Bern: Huber.
- *Garfinkel, L., Auerbach, O., & Joubert, L. (1985). Involuntary smoking and lung cancer: a case-control study. *Journal of the National Cancer Institute*, 75, 463-469.
- Garvey, W., D. & Griffith, B., C. (1971). Scientific Communication: Its role in the conduct of research and the creation of knowledge. *American Psychologist*, 26, 349-362.
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L., M. & Woloshin, S. (2008). Helping Doctors and Patients Make Sense of Health Statistics. *Psychological Science in the Public Interest*. 8(2). 53-96.
- Gilpin, A., R. (2008). r-equivalent, Meta-Analysis, and Robustness. An Empirical Examination of Rosenthal and Rubin's Effect Size Indicator. *Educational and Psychological Measurement*. 68(1), 42-57.
- Glass, G., V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 5, 3-8.
- Green, A. (1986). *On Private Madness*. London: H. Karnac (Books) Ltd.
- Greenland, S. & Longnecker, M., P. (1992). Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. *American Journal of Epidemiology*, 135, 1301-1309.
- Guzzo, R., A., Jackson, S., E. & Katzell, R., A. (1987). Meta-analysis analysis. *Research in Organizational Behavior*, 9, 407-442.
- Hackshaw, A., K. (1998). Lung cancer and passive smoking. *Statistical Methods in Medical Research*, 7, 119-136.

Reference

- Hackshaw, A., K., Law, M., R. & Wald, N., J. (1997). The accumulated evidence on lung cancer and environmental tobacco smoke. *British Medical Journal*, *315*, 980-988.
- Hayes, A., F. & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, *1*, 77-89.
- Hedges, L., V. (1982). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490-499.
- Hedges, L., V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
- Hedges L., V. (1987). Commentary on pooling the results of clinical trials. *Statistics in Medicine* *6*, 381-385
- Hedges, L., V. & Vevea, J., L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486-504.
- Howard, G., S., Maxwell, S., E. & Fleming, K., J. (2000). The Proof of the Pudding: Illustration of the Relative Strengths of Null Hypothesis, Meta-Analysis, and Bayesian Analysis. *Psychological Methods*. *5(3)*, 315-332.
- Huang, D., Guan, P., Shi, H., He, Q. & Zhoum B. (2008). Reliability and accuracy of interview data in non-smoking female lung cancer case-control study. *Journal of Experimental Clinical Cancer Research*, *27(1)*.
- Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York: Russel Sage.
- Hunter, J., E., Schmidt, F., L. & Jackson, G., B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.
- Hunter, J., E. & Schmidt, F., L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative knowledge in psychology. *International Journal of Selection and Assessment*, *8*, 275-292.
- Hunter, J., E. & Schmidt, F., L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2. Ed.). Thousand Oaks, CA: Sage Publications.
- Hunter, J., E., Schmidt, F., L. & Jackson, G., B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Reference

- *Johnson, K., C., Hu, J. & Mao, Y. (2001). Lifetime residential and workplace exposure to environmental tobacco smoke and lung cancer in never-smoking women, Canada. *International Journal of Cancer*, *93*, 902-906.
- *Kabat, G., C. & Wynder, E., L. (1984). Lung cancer in nonsmokers. *Cancer*, *53*, 1214-1221.
- *Kabat, G., C., Stellman, S., D. & Wynder, E., L. (1995). Relation between exposure to environmental tobacco smoke and lung cancer in lifetime nonsmokers. *American Journal of Epidemiology*, *142*, 141-148.
- *Kalandidi, A., Katsouyanni, K., Voropoulou, N., Bastas, G., Saracci, R. & Trichopoulos, D. (1990). Passive smoking and diet in the etiology of lung cancer among non-smokers. *Cancer Causes and Control*, *1*, 15-21.
- Koetse, M., J., Florax, R., J., G., M. & de Groot, H., L., F. (2007) The Impact of Effect Size Heterogeneity on Meta-Analysis: A Monte Carlo Experiment. Retrieved November 25th, 2008, from <http://ideas.repec.org/p/dgr/uvatin/20070052.html>.
- *Koo, L. C., Ho, J. H. C. & Saw, D. (1984). Is passive smoking an added risk factor for lung cancer in Chinese women? *Journal of Experimental & Clinical Cancer Research*, *3*, 277-284.
- *Kreuzer, M., Krauss, M., Kreienbrock, L., Jockel, K. H., & Wichmann, H. E. (2000). Environmental tobacco smoke and lung cancer: a case-control study in Germany. *American Journal of Epidemiology*, *151*, 241-250.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage.
- *Lee, C. H., Ko, Y. C., Goggins, W., Huangd, J. J., Huang, M. S., Kao, E. L., et al. Lifetime environmental exposure to tobacco smoke and primary lung cancer of non-smoking Taiwanese women. *International Journal of Epidemiology*, *29*, 224-231.
- *Lee, P. N., Chamberlain, J., & Alderson, M. R. (1986). Relationship of passive smoking to risk of lung cancer and other smoking-associated diseases. *British Journal of Cancer*, *54*, 97-105.

Reference

- Le Vois, M.E., & Layard, M. W. (1994). Inconsistency between workplace and spousal studies of environmental tobacco smoke and lung cancer. *Regulatory Toxicology and Pharmacology*, 19, 309-316.
- Lewis, S., Clarke, M., (2001). Forest plots: trying to see the wood and the trees. *British Medical Journal*; 322, 1479-1480
- Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.
- Lipsey, M. W. (1994). Identifying potentially interesting variables and analysis opportunities. In H. M. Cooper, & L. V. Hedges (Ed.), *The Handbook of Research Synthesis* (pp. 111-123). New York: Russell Sage.
- Lipsey, M. W., Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
- Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2004). *Practical resources for assessing and reporting intercoder reliability in content analysis research projects*. Retrieved November, 4, 2008 from http://www.slis.indiana.edu/faculty/hrosenba/www/Research/methods/lombard_reliability.pdf
- Mengersen, K. L., Tweedie, R. L., & Biggerstaff, B. (1995). The impact of method choice in meta-analysis. *Australian & New Zealand Journal of Statistics*, 37, 19-44.
- Morris, R., D. (1994). Meta-analysis in Cancer Epidemiology. *Environmental Health Perspectives*. 102(8), 61-66.
- Mundy, K.,M., & Stein, K., F. (2008). Meta-analysis as a basis for evidence-based practice: The question is, why not?. *Journal of the American Psychiatric Nurses Association*. 14(4), 326-328
- Orwin, R. G. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8, 157-159.
- Pahl, H. (2009). Reanalyse und Erweiterung einer Metaanalyse: Kennzeichnung der Subjektivität durch den Stellschraubenansatz. *Diplomarbeit [Thesis]*. Universität Mannheim.
- Pearson, K. (1904). Report on certain enteric fever inoculation statistics. *British Medical Journal*, 2, 1243-1246.

Reference

- Peshkin, A. (1988). In Search of Subjectivity - One's Own. *Educational Researcher*, 17, 17-21.
- Peterson, R. A., & Brown, S. P. (2005). On the use of beta coefficients in meta-analysis. *Journal of Applied Psychology*, 90, 175-181.
- Popper, K. R. (1982). *Die Logik der Forschung*. Tübingen: Mohr.
- Pron, G., E., Burch, J., D., Howe, G., R. & Miller, A., B. (1988). The Reliability of Passive Smoking Histories Reported in a Case-Control Study of Lung Cancer. *American Journal of Epidemiology*, 127(2), p 267-273.
- *Rapiti, E., Jindal, S. K., Gupta, D., & Boffetta, P. (1999). Passive smoking and lung cancer in Chandigarh, India. *Lung Cancer*, 23, 183-189.
- *Reynolds, P., von Behren, J., Fontham, E. T., Correa, P., Wu, A., Buffler, P. A. & Greenberg, R. S. (1996). Occupational exposure to environmental tobacco smoke. *The Journal of the American Medical Association*, 275, 441-442.
- Rosenberg, M. (2005). The file drawer-problem revisited: A general weighted method for calculating fail-safe numbers in meta-analysis. *Evolution*, 59, 464-468.
- Rosenthal, R. (1979). The "File drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Rosenthal, R., & Rubin, D. B. (1979). Comparing significance levels of independent studies. *Psychological Bulletin*, 86, 1165-1168.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Ed.). (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Chichester, England: Wiley.
- Rustenbach, S. J. (2003). *Metaanalyse. Eine anwendungsorientierte Einführung*. Bern, Switzerland: Verlag Hans Huber.
- Sánchez-Meca, J. & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: a Monte Carlo comparison of statistical power and Type I error. *Quality and Quantity*, 31(4), 385-399(15).

Reference

- Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.
- Schulze, R., (2004). *Meta Analysis, a Comparison of Approaches*. Hogrefe & Huber Publications.
- *Schwartz, A. G., Yang, P., & Swanson, G. M. (1996). Familial risk of lung cancer among nonsmokers and their relatives. *American Journal of Epidemiology, 144*, 554-562.
- Shannon H. (2008). A statistical note on Karl Pearson's 1904 metaanalysis. *The James Lind Library*. (www.jameslindlibrary.org).
- *Shimizu, H., Morishita, M., Mizuno, K., Masuda, T., Ogura, Y., Santo, M., et al. (1988). A casecontrol study of lung cancer in nonsmoking women. *The Tohoku Journal of Experimental Medicine, 154*, 389-397.
- Slavin, R. E. (1987). Best-evidence synthesis: Why less is more. *Educational Researcher, 16*, 15-16.
- Smith, M. L., Glass, G. V., (1977). Metaanalysis of psychotherapy outcome studies. *American Psychologist, 32*, 752-760.
- Spector, T., D. & Thomson, S.G. (1991). The potential and limitations of meta-analysis. *Journal of Epidemiology and Community Health, 45*, 89-92
- Stayner, L., Bena, J., Sasco, A. J., Smith, R., Steenland, K., Kreuzer, M., & Straif, K. (2007). Lung cancer risk and workplace exposure to environmental tobacco smoke. *American Journal of Public Health, 97*, 545-551.
- Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on a choice of axis. *Journal of Clinical Epidemiology, 54*, 1046-1055.
- Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., et al. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *The Journal of the American Medical Association, 283*, 2008-2012.

Reference

-
- *Sun, X.-W., Dai, X.-D., Lin, C.-Y., Shi, Y.-B., Ma, Y.-Y., & Li, W. (1996). Passive smoking and lung cancer among nonsmoking women in Harbin, China. International symposium on lifestyle factors and human lung cancer, China 1994. *Lung Cancer, 14*, p. 237.
- Sutton, A. J., & Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine, 27*, 625-650.
- Vandenbroucke, J., P., Elm, E., Altman, D., G., Gøtzsche, P., C., Mulrow, C., D., Pocock, S., J., Poole, C. Schlesselman, J., J. & Egger, M. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Explanation and Elaboration. *Epidemiology, 18 (6)*, 805-835.
- Viechtbauer, W. (2007a). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie [Journal of Psychology], 215*, 104-121.
- Viechtbauer, W. (2007b). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology, 60*, 29-60.
- *Wang, L., Lubin, J. H., Zhang, S. R., Metayer, C., Xia, Y., Brenner, A., et al. (2000). Lung cancer and environmental tobacco smoke in a non-industrial area of China. *International Journal of Cancer, 88*, 139-154.
- Wang, M. C., & Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods, 3*, 46-54.
- *Wang, T.-J., Zhou, B.-S., & Shi, J.-P. (1996). Lung cancer in nonsmoking Chinese women: a case-control study. International symposium on lifestyle factors and human lung cancer, China 1996. *Lung Cancer, 14*: S93-S98.
- Wanous, J. P., Sullivan, S. E., & Mailnak, J. (1989). The role of judgment calls in meta-analysis. *Journal of Applied Psychology, 74*, 259-264.
- Wells A., J. (1998). Lung cancer from passive smoking at work. *American Journal of Public Health, 88*, 1025-1029.
- White, H. D. (1994). Scientific communication and literature retrieval. In H. Cooper, & L. V. Hedges (Ed.), *The Handbook of Research Synthesis* (pp. 41-55). New York, NY: Russel Sage.
- Wilson SPSS Macro. From <http://mason.gmu.edu/~dwilsonb/ma.html>.

Reference

- Wittmann, W. W. (1985). *Evaluationsforschung*. Berlin: Springer.
- Wittmann, W. W., & Matt, G. E. (1986). Meta-Analyse als Integration von Forschungsergebnissen am Beispiel deutschsprachiger Arbeiten zur Effektivität von Psychotherapie. *Psychologische Rundschau*, 37, 20-40.
- Wittmann, W. W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology*, 2nd ed. (pp. 505-560). New York: Plenum Press.
- Wortman, P. M. (1994). Judging research quality. In H. Cooper, & L. V. Hedges (Ed.), *The Handbook of Research Synthesis* (pp. 97-109). New York, NY: Russell Sage.
- *Wu, A. H., Henderson, B. E., Pike, M. C., & Yu, M. C. (1985). Smoking and other risk factors for lung cancer in women. *Journal of the National Cancer Institute*, 74, 747-751.
- *Wu-Williams, A. H., Dai, X. D., Blot, W., Xu, Z. Y., Sun, X. W., Xiao, H. P., et al. (1990). Lung cancer among women in north-east China. *British Journal of Cancer*, 62, 982-987.
- Zack, M. (1996). SAS Macro to calculate trend estimation based on the Greenland and Longnecker reference. Received by E-mail from the author.
- *Zaridze, D., Maximovitch, D., Zemlyanaya, G., Aitakov, Z. N., & Boffetta, P. (1998). Exposure to environmental tobacco smoke and risk of lung cancer in non-smoking women from Moscow, Russia. *International Journal of Cancer*, 75, 335-338.
- *Zhong, L., Goldberg, M. S., Gao, Y. T., & Jin, F. (1999). A case-control study of lung cancer and environmental tobacco smoke among nonsmoking women living in Shanghai, China. *Cancer Causes Control*, 10, 607-61.
- Ziegler, A., Lange, S., Bender, R., (2004). Systematische Übersichten und Meta Analysen. *Dtsch. Med. Wochenschr*, 129, 11-15.
- Ziman, J. (1969). Information, communication, knowledge. *Nature*, 224, 318-324.

8 Appendix

What is being presented here is the index. You will find the entire appendix on the enclosed CD. If you have any questions feel free to contact the author: robin.wuerfel@gmail.com.

Table 6 Appendix Content

Section	Content
A. Data and outputs of the first replication phase	<ul style="list-style-type: none"> a. Mean Effect and moderator analyses of the key study design features b. Exposure response analyses c. Sensitivity analyses d. Funnel Plot
B. Data and outputs of the second replication phase	<ul style="list-style-type: none"> a. Mean Effect and moderator analyses of the key study design features b. Exposure response analyses c. Sensitivity analyses d. Funnel Plot
C. Data and outputs CRM	<ul style="list-style-type: none"> a. CRM theory one to four b. Bad guy analysis c. Forest plot
D. Data and outputs advanced analyses	<ul style="list-style-type: none"> a. Quality moderator calculations via block two and four b. Artefact corrections
E. Original studies	<ul style="list-style-type: none"> a. Thesis as PDF b. Stayner et al. (2007) c. Primary studies d. Pahl (2009).pdf
F. Materials	<ul style="list-style-type: none"> a. List of used software b. Documentation of macros and routines c. Documentation of R 2.6.1 source code
G. Data	<ul style="list-style-type: none"> a. Coding manual b. Coding sheet c. Interrater reliability

Eidesstattliche Erklärung

Ich versichere, dass ich die beiliegende Diplomarbeit ohne Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt und die den benutzten Quellen wörtlich oder inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Diese Arbeit hat noch keiner Prüfungsbehörde vorgelegen.

Ort, Datum, Unterschrift

Freiwillige Erklärung

Ich stimme zu, dass meine durch Prof. Dr. Michael Bosnjak betreute Diplomarbeit mit dem Titel „Reanalyse und Erweiterung einer Metaanalyse: Kennzeichnung der Subjektivität durch den Stellschraubenansatz“ nach Beendigung der Diplomprüfung zu wissenschaftlichen Zwecken im Bereich der Fakultätsbibliothek für Sozialwissenschaften aufgestellt und zugänglich gemacht wird (Veröffentlichungen nach §§ 6 Abs. 1 UrhG), und hieraus im Rahmen des § 51 UrhG zitiert werden kann.

Sämtliche Verwertungsrechte nach § 15 UrhG verbleiben beim Verfasser der Diplomarbeit.

Ort, Datum, Unterschrift